



Université  
de Toulouse

# THÈSE

En vue de l'obtention du

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Genie industriel

---

Présentée et soutenue par :

Mme PAULA ANDREA POTES RUIZ

le lundi 24 novembre 2014

Titre :

GENERATION DE CONNAISSANCES A L'AIDE DU RETOUR  
D'EXPERIENCE: APPLICATION A LA MAINTENANCE INDUSTRIELLE

---

Ecole doctorale :

Systèmes (Systèmes)

Unité de recherche :

Laboratoire de Génie de Productions de l'ENIT (E.N.I.T-L.G.P.)

Directeur(s) de Thèse :

M. BERNARD GRABOT

Rapporteurs :

M. BENOÎT IUNG, UNIVERSITE DE LORRAINE

M. ENGELBERT MEPHU NGUIFO, UNIVERSITE CLERMONT-FERRAND 2

Membre(s) du jury :

Mme PASCALE ZARATE, UNIVERSITE TOULOUSE 1, Président

M. BERNARD GRABOT, ECOLE NATIONALE D'INGENIEUR DE TARBES, Membre

M. BERNARD KAMU-FOGUEM, ECOLE NATIONALE D'INGENIEUR DE TARBES, Membre

M. BRUNO VALLESPER, UNIVERSITE BORDEAUX 1, Membre



*A mis padres,  
por ser el pilar fundamental  
en todo lo que soy*

***La vida no es la que uno vivió, sino la que uno  
recuerda, y cómo la recuerda para contarla.***

Gabriel García Márquez



# REMERCIEMENTS

Ces travaux de recherche ont été effectués au sein de l'équipe Systèmes Décisionnels et Cognitifs (SDC) du Laboratoire Génie de Production (LGP) de l'Ecole Nationale d'Ingénieurs de Tarbes (ENIT). Je souhaite remercier toutes les personnes qui ont contribué à la réalisation de cette thèse.

Je tiens tout d'abord à remercier mon directeur de thèse, M. Bernard Grabot, Professeur à l'ENIT, et mon co-encadrant, M. Bernard Kamsu-Foguem, Maître de conférences à l'ENIT, pour la confiance qu'ils m'ont accordée en acceptant d'encadrer mon travail de recherche. J'ai particulièrement apprécié travailler avec eux, pour leurs qualités humaines et scientifiques, leur bonne humeur permanente, leurs conseils et leurs encouragements, qui m'ont aidée à mener à bien ce travail.

J'associe également à ces remerciements, M. Daniel Noyes, ancien directeur du Laboratoire Génie de Production, et M. Laurent Geneste, chef de l'équipe SDC, pour leur accueil, leur disponibilité et leur gentillesse, ainsi que tous les membres du laboratoire pour leurs échanges aussi bien humains que scientifiques.

J'adresse aussi mes remerciements aux personnes qui ont accepté de participer à mon jury de thèse : Mme. Pascale Zaraté, Professeur à l'Université Toulouse 1, pour avoir accepté de présider ce jury ainsi que M. Benoît lung, Professeur à l'Université de Lorraine, et M. Engelbert Mephu-Nguifo, Professeur à l'Université Clermont Ferrand 2, pour avoir accepté d'être rapporteurs de mes travaux. J'associe à ces remerciements M. Bruno Vallespir, Professeur à l'Université Bordeaux, pour avoir accepté de participer à ce jury en tant qu'examineur.

De façon plus personnelle, je tiens à remercier chaleureusement toutes les personnes qui, de près ou de loin, m'ont accompagnée au cours de ces années et plus particulièrement mes collègues doctorants (de différentes générations) avec lesquels j'ai partagé des moments agréables et à qui je souhaite les meilleures choses. Merci à Memon, Simon, Floran, Damien, Stéphanie, Alberto, Olivier, Corentin, Christian, Malik, Adrien, Romain et tous les autres, pour leur amabilité et leurs encouragements. Merci plus particulièrement à Karima, Samira et Yasmina avec qui j'ai partagé de très bons moments et qui ont toujours été présentes pour m'encourager dans les périodes difficiles. J'adresse également mes sincères remerciements à Éric pour sa gentillesse, ses conseils et son soutien pendant ces années.

Comment ne pas saisir l'occasion de remercier mes compatriotes Vanessa, Victor et Juan Camilo pour leur amitié sans faille et leurs encouragements. Merci particulièrement à Juan Diego pour son amitié et son soutien depuis le début de cette aventure en France.

Je ne saurais terminer ces remerciements sans citer mes proches. Un grand merci à Pierre et sa famille pour m'avoir accueillie, soutenue et encouragée tout au long de ces travaux. Et enfin, je ne peux pas oublier mes parents, Alvaro et Olivia, ainsi que mon frère Alvaro et ma cousine Naty, qui m'ont tant manqué durant ces années, mais dont leurs encouragements et leur présence constante malgré la distance, m'ont donné la force d'aller jusqu'à la fin. Merci à vous !

***La connaissance s'acquiert par l'expérience, tout le reste n'est que de l'information.***

Albert Einstein



# RESUME

Les travaux de recherche présentés dans ce mémoire s'inscrivent dans le cadre de la valorisation des connaissances issues des expériences passées afin d'améliorer les performances des processus industriels.

La connaissance est considérée aujourd'hui comme une ressource stratégique importante pouvant apporter un avantage concurrentiel décisif aux organisations. La gestion des connaissances (et en particulier le retour d'expérience) permet de préserver et de valoriser des informations liées aux activités d'une entreprise afin d'aider la prise de décision et de créer de nouvelles connaissances à partir du patrimoine immatériel de l'organisation. Dans ce contexte, les progrès des technologies de l'information et de la communication jouent un rôle essentiel dans la collecte et la gestion des connaissances. L'implémentation généralisée des systèmes d'information industriels, tels que les ERP (*Enterprise Resource Planning*), rend en effet disponible un grand volume d'informations issues des événements ou des faits passés, dont la réutilisation devient un enjeu majeur. Toutefois, ces fragments de connaissances (les expériences passées) sont très contextualisés et nécessitent des méthodologies bien précises pour être généralisés.

Etant donné le potentiel des informations recueillies dans les entreprises en tant que source de nouvelles connaissances, nous proposons dans ce travail une démarche originale permettant de générer de nouvelles connaissances tirées de l'analyse des expériences passées, en nous appuyant sur la complémentarité de deux courants scientifiques : la démarche de Retour d'Expérience (REx) et les techniques d'Extraction de Connaissances à partir de Données (ECD).

Le couplage REx-ECD proposé porte principalement sur : *i*) la modélisation des expériences recueillies à l'aide d'un formalisme de représentation de connaissances afin de faciliter leur future exploitation, et *ii*) l'application de techniques relatives à la fouille de données (ou *data mining*) afin d'extraire des expériences de nouvelles connaissances sous la forme de règles. Ces règles doivent nécessairement être évaluées et validées par les experts du domaine avant leur réutilisation et/ou leur intégration dans le système industriel. Tout au long de cette démarche, nous avons donné une place privilégiée aux Graphes Conceptuels (GCs), formalisme de représentation des connaissances choisi pour faciliter le stockage, le traitement et la compréhension des connaissances extraites par l'utilisateur, en vue d'une exploitation future.

Ce mémoire s'articule en quatre chapitres. Le premier constitue un état de l'art abordant les généralités des deux courants scientifiques qui contribuent à notre proposition : le REx et les techniques d'ECD. Le second chapitre présente la démarche REx-ECD proposée, ainsi que les outils mis en œuvre pour la génération de nouvelles connaissances afin de valoriser les informations disponibles décrivant les expériences passées. Le troisième chapitre présente une méthodologie structurée pour interpréter et évaluer l'intérêt des connaissances extraites lors de la phase de post-traitement du processus d'ECD. Finalement, le dernier chapitre expose des cas réels d'application de la démarche proposée à des interventions de maintenance industrielle.

**Mots clés :** Retour d'Expérience, Extraction de Connaissances à partir de Données, Formalisation des connaissances, Extraction de règles d'association, Maintenance industrielle





# ABSTRACT

The research work presented in this thesis relates to knowledge extraction from past experiences in order to improve the performance of industrial process.

Knowledge is nowadays considered as an important strategic resource providing a decisive competitive advantage to organizations. Knowledge management (especially the experience feedback) is used to preserve and enhance the information related to a company's activities in order to support decision-making and create new knowledge from the intangible heritage of the organization. In that context, advances in information and communication technologies play an essential role for gathering and processing knowledge. The generalised implementation of industrial information systems such as ERPs (Enterprise Resource Planning) make available a large amount of data related to past events or historical facts, which reuse is becoming a major issue. However, these fragments of knowledge (past experiences) are highly contextualized and require specific methodologies for being generalized.

Taking into account the great potential of the information collected in companies as a source of new knowledge, we suggest in this work an original approach to generate new knowledge based on the analysis of past experiences, taking into account the complementarity of two scientific threads: Experience Feedback (EF) and Knowledge Discovery techniques from Databases (KDD).

The suggested EF-KDD combination focuses mainly on: *i)* modelling the experiences collected using a knowledge representation formalism in order to facilitate their future exploitation, and *ii)* applying techniques related to data mining in order to extract new knowledge in the form of rules. These rules must necessarily be evaluated and validated by experts of the industrial domain before their reuse and/or integration into the industrial system. Throughout this approach, we have given a privileged position to Conceptual Graphs (CGs), knowledge representation formalism chosen in order to facilitate the storage, processing and understanding of the extracted knowledge by the user for future exploitation.

This thesis is divided into four chapters. The first chapter is a state of the art addressing the generalities of the two scientific threads that contribute to our proposal: EF and KDD. The second chapter presents the EF-KDD suggested approach and the tools used for the generation of new knowledge, in order to exploit the available information describing past experiences. The third chapter suggests a structured methodology for interpreting and evaluating the usefulness of the extracted knowledge during the post-processing phase in the KDD process. Finally, the last chapter discusses real case studies dealing with the industrial maintenance domain, on which the proposed approach has been applied.

**Keywords:** Experience feedback, Knowledge discovery from databases, Knowledge formalization, Association rule mining, Industrial maintenance.



# TABLE DES MATIERES

<b>RESUME .....</b>	<b>I</b>
<b>ABSTRACT.....</b>	<b>III</b>
<b>INTRODUCTION GENERALE .....</b>	<b>1</b>
<b>I EXPLOITER LES EXPERIENCES PASSES .....</b>	<b>5</b>
I.1 LA CONNAISSANCE ET LA GESTION DES CONNAISSANCES : DEUX ENJEUX IMPORTANTS POUR LES ENTREPRISES .....	6
I.1.1 NOTION DE CONNAISSANCE .....	7
I.1.2 TYPES DE CONNAISSANCE .....	7
I.1.3 MODES DE CREATION ET DE TRANSFERT DES CONNAISSANCES .....	9
I.1.4 LA GESTION DES CONNAISSANCES AU SEIN DES ORGANISATIONS .....	10
I.2 GESTION DES EXPERIENCES .....	12
I.2.1 DEFINITION D'UNE EXPERIENCE .....	13
I.2.2 APPROCHES DE GESTION DES EXPERIENCES .....	14
I.2.3 LE RETOUR D'EXPERIENCE .....	14
I.2.4 TYPES DE RETOUR D'EXPERIENCE .....	15
I.3 MISE EN ŒUVRE D'UN SYSTEME DE RETOUR D'EXPERIENCE .....	16
I.3.1 DIMENSION « EXPERIENCES ET CONNAISSANCES » .....	17
I.3.2 DIMENSION « ACTIVITES ET PROCESSUS » .....	17
I.3.3 DIMENSION « ACTEURS ET COMPETENCES » .....	17
I.3.4 DIMENSION « TECHNIQUES ET OUTILS » .....	17
I.4 FORMALISATION DES CONNAISSANCES .....	18
I.4.1 APPROCHES BASEES SUR LES FRAMES .....	19
I.4.2 APPROCHES BASEES SUR LES GRAPHS .....	19
I.5 L'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES.....	20
I.5.1 DEFINITION DU PROCESSUS D'ECD .....	21

---

I.5.2	PRINCIPALES ETAPES D'UN PROCESSUS D'ECD .....	22
I.5.3	EXTRACTION DE REGLES D'ASSOCIATION .....	26
I.6	BILAN .....	34
<b>II</b>	<b>COUPLAGE REX-ECD : UNE DEMARCHE ORIGINALE DE GESTION DES CONNAISSANCES .....</b>	<b>37</b>
II.1	POSITIONNEMENT DE LA DEMARCHE DE RECHERCHE .....	38
II.1.1	POSITIONNEMENT .....	38
II.1.2	INTERETS D'UN COUPLAGE REX-ECD .....	40
II.2	FORMALISATION DES CONNAISSANCES A L'AIDE DES GRAPHS CONCEPTUELS.....	41
II.2.1	GENERALITES .....	42
II.2.2	VOCABULAIRE DU DOMAINE.....	42
II.2.3	NOTION DE GRAPHE CONCEPTUEL .....	44
II.2.4	SUBSOMPTION ET HOMOMORPHISME : DES OUTILS DE RAISONNEMENT .....	44
II.2.5	GRAPHS CONCEPTUELS EMBOITES .....	46
II.2.6	INTERET DES GCs DANS NOTRE APPROCHE REX-ECD.....	46
II.3	DESCRIPTION DE LA DEMARCHE GLOBALE REX-ECD.....	47
II.3.1	PRETRAITEMENT DE L'ECD ASSOCIE A LA PHASE DE CAPITALISATION DU REX .....	48
II.3.2	TECHNIQUES DE DATA MINING ASSOCIEES A LA PHASE DE TRAITEMENT DU REX.....	50
II.3.3	POST-TRAITEMENT DE L'ECD ASSOCIE A LA PHASE D'EXPLOITATION DU REX.....	59
II.4	BILAN .....	62
<b>III</b>	<b>UNE DEMARCHE INTERACTIVE POUR LE POST-TRAITEMENT DANS LE PROCESSUS D'ECD.....</b>	<b>63</b>
III.1	ROLE DE L'EXPERT DU DOMAINE DANS LE PROCESSUS D'ECD .....	64
III.2	INTERACTION ET COOPERATION ENTRE L'EXPERT DU DOMAINE ET L'EXPERT D'ECD.....	65
III.3	IMPORTANCE DES ASPECTS D'INTERACTION HOMME-MACHINE DANS LE PROCESSUS.....	67
III.4	INTERPRETATION ET EVALUATION DES CONNAISSANCES EXTRAITES.....	68
III.4.1	EVALUATION OBJECTIVE : UNE ANALYSE D'UN POINT DE VUE STATISTIQUE.....	69
III.4.2	EVALUATION SEMANTIQUE : INTERPRETATION ET VISUALISATION DES RESULTATS....	72
III.4.3	EVALUATION SUBJECTIVE : RECHERCHE DES TYPES SPECIFIQUES DE REGLES.....	82
III.5	BILAN .....	87
<b>IV</b>	<b>GESTION DES CONNAISSANCES EN MAINTENANCE INDUSTRIELLE .....</b>	<b>89</b>
IV.1	LA MAINTENANCE INDUSTRIELLE : UN PROCESSUS STRATEGIQUE DANS LES ENTREPRISES..	91

---

---

IV.2 SYSTEMES DE MAINTENANCE BASES SUR LA CONNAISSANCE .....	92
IV.3 CAS D'APPLICATION.....	95
IV.3.1 GENERALITES .....	95
IV.3.2 BASES DE DONNEES CONSIDEREES .....	95
IV.3.3 DU RAPPORT D'INTERVENTION A L'EXPERIENCE FORMALISEE .....	99
IV.3.4 DU RAPPORT D'INTERVENTION A LA CONNAISSANCE.....	104
IV.4 BILAN .....	130
<b>CONCLUSION ET PERSPECTIVES .....</b>	<b>133</b>
<b>TABLE DE FIGURES .....</b>	<b>139</b>
<b>LISTE DE TABLEAUX .....</b>	<b>141</b>
<b>ALGORITHMES .....</b>	<b>143</b>
<b>BIBLIOGRAPHIE .....</b>	<b>145</b>
<b>BIBLIOGRAPHIE DE L'AUTEUR .....</b>	<b>161</b>
<b>ANNEXES .....</b>	<b>163</b>



# INTRODUCTION GENERALE

---

La connaissance est de plus en plus considérée aujourd'hui comme une ressource immatérielle stratégique des organisations, pouvant leur apporter un avantage concurrentiel décisif. En effet, l'économie mondiale est en train de passer d'un système de valeur basé sur la production matérielle à un système basé sur les connaissances (Chen, 2010) afin d'améliorer les performances des processus industriels. Cet état de fait a débouché sur le concept d' « entreprise basée sur la connaissance ». La gestion des connaissances est ainsi devenue un enjeu très important dans de nombreux domaines industriels ; elle doit rendre possible un processus de création de valeur à partir des différentes sources de connaissances de l'organisation, qui peuvent être disponibles sous des formes distinctes : dans l'esprit des employés, dans les procédures de travail, les normes, les documents internes, les bases de données, etc. (Bergmann, 2002).

Dans ce contexte, les progrès des technologies de l'information et de la communication (incluant les technologies de stockage) jouent un rôle essentiel dans la collecte et la gestion des connaissances. L'implémentation généralisée des systèmes d'information industriels tels que les ERP (*Enterprise Resource Planning*) rend en effet disponible un grand volume d'informations issues des processus industriels, dont la réutilisation devient un enjeu majeur.

Dans nos travaux de recherche, une attention particulière a été portée au domaine de la maintenance industrielle, dans lequel une grande quantité d'informations est collectée quotidiennement (Benomrane et al., 2013). Aujourd'hui, même des opérations ou procédures de maintenance simples sont enregistrées dans les systèmes d'information, tels que les outils de GMAO (Gestion de Maintenance Assistée par Ordinateur). Or, la gestion des connaissances en maintenance apparaît de plus en plus comme un facteur essentiel afin d'améliorer les performances des entreprises, qui dépendent largement de la qualité de la maintenance de leurs ressources opérationnelles. Ainsi, la gestion des connaissances doit pouvoir permettre de préserver et de valoriser les connaissances issues de l'analyse des événements ou des faits passés en vue d'améliorer les plans de maintenance actuels, les stratégies de maintenance, le diagnostic des équipements, les mesures de prévention à mettre en œuvre et/ou les procédures et actions menées lors des interventions.

Au niveau industriel, il existe donc un besoin réel de capitaliser et d'exploiter les informations issues des expériences passées afin d'améliorer la prise de décision et de créer de nouvelles connaissances à partir du patrimoine immatériel des organisations. Les approches de Retour d'Expérience (REx) sont ainsi devenues une nécessité stratégique pour les entreprises qui commencent à accorder plus d'attention aux expériences acquises au cours des activités antérieures. Toutefois, ces connaissances spécifiques - ou fragments de connaissances - (les expériences passées) sont souvent très contextualisées et nécessitent des méthodologies complexes pour être généralisées.

Les grandes quantités d'information recueillies dans les systèmes d'information industriels ont un fort potentiel pour fournir de nouvelles connaissances. Une manière de généraliser les expériences passées dans les démarches classiques de retour d'expérience est la production de règles expertes et de standards par des comités d'experts de différents domaines. Toutefois, ces approches présentent des inconvénients, par exemple le manque de disponibilité des experts pour analyser les expériences passées de l'entreprise, la difficulté d'évaluer la pertinence des nouvelles règles ou connaissances générées, ou l'incapacité des experts à analyser les volumes importants d'informations contenues dans les bases de données. Les techniques d'Extraction de Connaissances à partir de Données (ECD) ont ainsi reçu une attention considérable de la part des chercheurs en sciences de l'information comme des décideurs au sein des entreprises. Leur enjeu est de valoriser ces bases de données, dont la taille croît de manière exponentielle, par des techniques complémentaires à la démarche de retour d'expérience. Ces techniques sont devenues essentielles dans de nombreux domaines industriels, incluant la conception de produits et de processus, la planification des ressources, le contrôle qualité, la maintenance des équipements, la gestion de la relation client, entre autres.

Malgré la complémentarité des deux approches émergentes REx et ECD, peu de travaux de recherche dans le domaine du retour d'expérience se sont intéressés à l'ECD comme à un moyen de générer des connaissances plus généralisées que les expériences. A cette fin, nous proposons dans ce mémoire une démarche de retour d'expérience intégrant des techniques de fouille de données, connues en anglais sous la dénomination de *data mining*. Nous suggérons la mise en œuvre d'un couplage REx-ECD portant principalement sur : *i)* la modélisation des expériences recueillies à l'aide d'un formalisme de représentation de connaissances afin de faciliter leur future exploitation, et *ii)* l'application de techniques de *data mining* afin d'extraire des expériences de nouvelles connaissances sous la forme de règles. Ces règles, générées à l'aide des algorithmes de fouille de données, doivent à notre avis nécessairement être évaluées et validées par les experts du domaine avant leur réutilisation et/ou leur intégration dans les systèmes industriels. Cette validation est ainsi une étape fondamentale de la démarche. De plus, tout au long de la démarche conceptuelle proposée, un intérêt particulier est accordé aux Graphes Conceptuels (GCs), formalisme de représentation de connaissances proposé pour faciliter la structuration sémantique, le traitement et la compréhension des connaissances par l'utilisateur, en vue d'une exploitation future.

Ce mémoire de thèse est structuré en quatre chapitres :

Le **Chapitre I**, intitulé « *Exploiter les expériences passées* », analyse l'état de l'art concernant les généralités des deux courants scientifiques qui contribuent à notre proposition : la démarche de Retour d'Expérience (REx) et les techniques d'Extraction de Connaissances à partir de Données (ECD). Nous présentons tout d'abord la gestion des connaissances et la gestion des expériences dans les



entreprises, et notamment le processus de retour d'expérience qui sert de base à notre proposition. Nous exposons ensuite les principaux formalismes de représentation de connaissances susceptibles d'être utilisés dans ce contexte d'étude. Pour finir, un état de l'art des différentes étapes d'un processus d'extraction de connaissances à partir de données est présenté, ainsi que les différentes techniques relatives à la fouille de données permettant d'analyser les bases de données disponibles afin d'extraire de nouvelles connaissances.

Le **Chapitre II**, « *Couplage REx-ECD : une démarche originale de gestion des connaissances* », s'articule autour d'un couplage entre un processus de retour d'expérience et un processus d'extraction de connaissances, afin de créer une mémoire d'entreprise (appelée dans ce mémoire « base de REx-ECD ») incluant différents types de connaissances issues des informations disponibles (expériences formalisées et règles extraites) (Figure A).

Nous présentons dans un premier temps le positionnement de la démarche conceptuelle proposée. Ensuite, nous exposons en détail le formalisme de représentation de connaissances suggéré pour supporter cette méthodologie. Finalement, après une description des grandes lignes de l'approche, nous présentons en détail les étapes de la démarche de retour d'expérience (i.e. capitalisation, traitement et exploitation), intégrant un processus de fouille de données (i.e. prétraitement, *data mining* et post-traitement) afin de modéliser les expériences recueillies et d'extraire de nouvelles connaissances pour améliorer la prise de décisions stratégiques. Nous montrerons dans ce chapitre la contribution de notre approche à la réutilisation des expériences passées, ainsi qu'à l'enrichissement du référentiel d'entreprise (ou référentiel métier) pour améliorer les processus industriels.

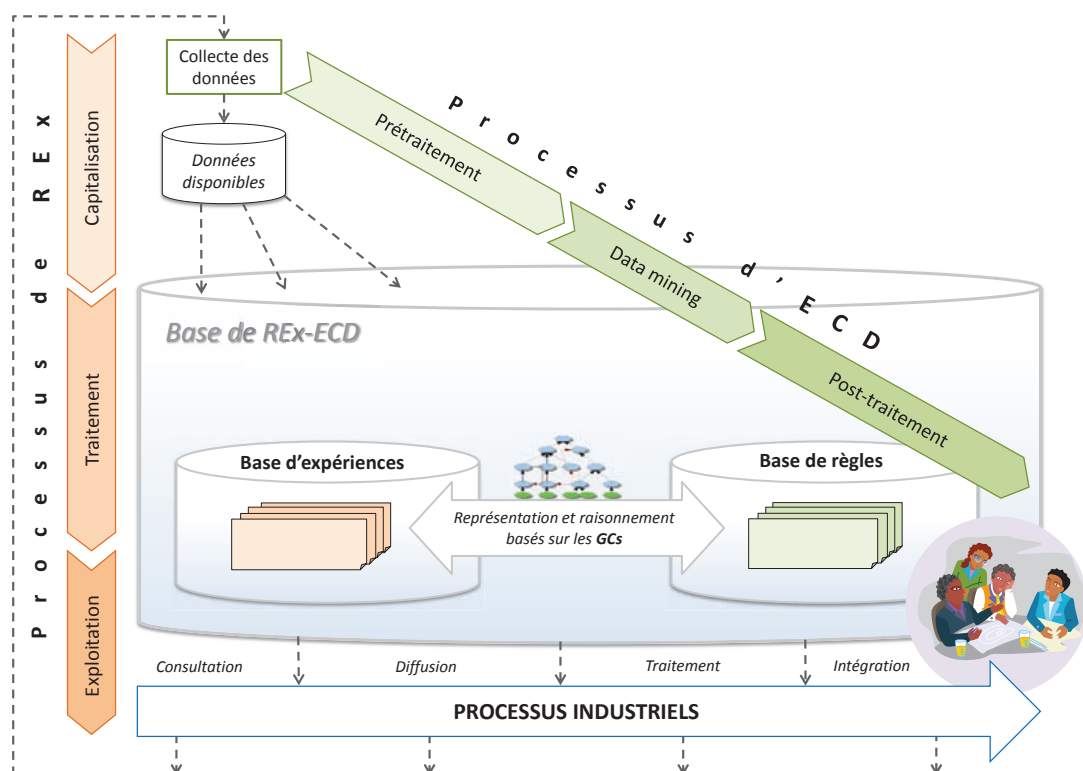


Figure A. Schéma général de la démarche conceptuelle REx-ECD

Le **Chapitre III**, intitulé « *Une démarche interactive pour le post-traitement dans le processus d'ECD* », est consacré à la dernière phase du processus d'extraction de connaissances à partir de données, abordée dans le Chapitre II : il s'agit d'analyser et d'évaluer les nouvelles connaissances extraites, et plus particulièrement les règles d'association (type de connaissances extraites sur lequel porte notre travail). Nous présentons une méthodologie structurée permettant d'interpréter, visualiser et filtrer les règles obtenues. Cette méthodologie favorise la collaboration entre experts de deux domaines différents : experts d'extraction de connaissances et experts métier, spécialistes du domaine d'application.

Dans ce contexte, nous présentons une démarche originale qui prend en compte trois manières d'évaluer et d'analyser les règles d'association : une évaluation objective afin d'interpréter les mesures d'intérêt associées aux règles, une évaluation sémantique pour analyser l'utilité de chaque règle et visualiser les relations existantes entre l'ensemble de règles obtenues, et une évaluation subjective basée sur un mécanisme de requête-réponses en vue de permettre à l'utilisateur de filtrer les règles.

Finalement, le **Chapitre IV**, intitulé « *Gestion des connaissances en maintenance industrielle* », aborde le contexte d'application qui sert de cadre à nos travaux de recherche et expose des cas d'étude permettant d'instancier la démarche proposée dans les chapitres II et III à des interventions de maintenance industrielle. Dans un premier temps, nous présentons le contexte de l'étude et proposons un état de l'art sur les systèmes d'aide à la maintenance basés sur les connaissances, en insistant sur leurs avantages, inconvénients et différences par rapport à notre approche générale. Puis, la deuxième partie de ce chapitre est consacrée à l'application de la démarche à des données réelles issues des industries aéronautique et pharmaceutique. Nous analysons les différents types de bases de données exploitées et les résultats obtenus.

La conclusion nous permettra de tirer un premier bilan sur cette étude, et de dresser des perspectives pour l'avenir de ce domaine qui nous semble très prometteur.

# I EXPLOITER LES EXPERIENCES PASSEES

---

I.1	LA CONNAISSANCE ET LA GESTION DES CONNAISSANCES : DEUX ENJEUX IMPORTANTS POUR LES ENTREPRISES .....	6
I.1.1	NOTION DE CONNAISSANCE .....	7
I.1.2	TYPES DE CONNAISSANCE .....	7
I.1.3	MODES DE CREATION ET DE TRANSFERT DES CONNAISSANCES .....	9
I.1.4	LA GESTION DES CONNAISSANCES AU SEIN DES ORGANISATIONS .....	10
I.2	GESTION DES EXPERIENCES .....	12
I.2.1	DEFINITION D'UNE EXPERIENCE .....	13
I.2.2	APPROCHES DE GESTION DES EXPERIENCES .....	14
I.2.3	LE RETOUR D'EXPERIENCE .....	14
I.2.4	TYPES DE RETOUR D'EXPERIENCE .....	15
I.3	MISE EN ŒUVRE D'UN SYSTEME DE RETOUR D'EXPERIENCE .....	16
I.3.1	DIMENSION « EXPERIENCES ET CONNAISSANCES » .....	17
I.3.2	DIMENSION « ACTIVITES ET PROCESSUS » .....	17
I.3.3	DIMENSION « ACTEURS ET COMPETENCES » .....	17
I.3.4	DIMENSION « TECHNIQUES ET OUTILS » .....	17
I.4	FORMALISATION DES CONNAISSANCES .....	18
I.4.1	APPROCHES BASEES SUR LES FRAMES .....	19
I.4.2	APPROCHES BASEES SUR LES GRAPHES .....	19
I.5	L'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES .....	20
I.5.1	DEFINITION DU PROCESSUS D'ECD .....	21

I.5.2	PRINCIPALES ETAPES D'UN PROCESSUS D'ECD .....	22
I.5.3	EXTRACTION DE REGLES D'ASSOCIATION .....	26
I.6	BILAN .....	34

Dans ce premier chapitre, nous présentons un état de l'art autour de deux approches générales qui contribuent à notre proposition. D'une part, la gestion des connaissances/expériences et les principaux formalismes de représentation de connaissances susceptibles d'être utilisés dans notre travail ; d'autre part, l'extraction de connaissances à partir de données et les différentes techniques relatives à la fouille de données (appelées techniques de *data mining* par la suite).

Plus particulièrement, nous décrivons dans un premier temps l'intérêt de la gestion des connaissances au sein des organisations et les différentes approches proposées dans la littérature favorisant la création et le transfert de connaissances dans les entreprises. Ensuite, nous présentons les généralités des approches de gestion des expériences dans un contexte industriel, et plus particulièrement la démarche de retour d'expérience qui sert de base à notre proposition. Finalement, nous présentons le processus d'extraction de connaissances à partir de données, ses principales étapes, ainsi que la technique de *data mining* sur lequel porte notre travail : l'extraction de règles d'association.

## **I.1 LA CONNAISSANCE ET LA GESTION DES CONNAISSANCES : DEUX ENJEUX IMPORTANTS POUR LES ENTREPRISES**

La gestion des connaissances (*knowledge management* en anglais) peut être définie comme un processus de création de valeur à partir des actifs intangibles d'une organisation (Lesser & Prusak, 2004). Ce processus implique notamment des notions de plusieurs domaines différents, tels que le comportement organisationnel, la gestion des ressources humaines, l'intelligence artificielle et les technologies de l'information et de la communication (TIC) (Liebowitz, 2001).

Le principal objectif des différentes approches basées sur la gestion des connaissances est de supporter la prise de décision et de créer de nouvelles connaissances à partir du patrimoine immatériel de l'entreprise pour structurer et échanger ces connaissances au moyen d'une « Communauté de Pratique » (Kimble & Bourdon, 2008). En effet, de nombreuses organisations commencent à considérer ces communautés de pratique comme des outils de collaboration, permettant d'encourager le partage des connaissances organisationnelles (Kimble & Bourdon, 2008).

Dans cette première partie du chapitre, nous allons aborder un état de l'art concernant les généralités sur la gestion des connaissances/expériences dans les entreprises afin de positionner nos travaux. Nous exposons ainsi quelques définitions associées : connaissance, types de connaissances, gestion des connaissances, expérience, gestion des expériences, retour d'expérience, etc.

### I.1.1 NOTION DE CONNAISSANCE

Plusieurs définitions de la connaissance sont proposées dans la littérature. Selon la norme AFNOR, la connaissance peut être définie comme : « *un ensemble de représentations, idées ou perceptions acquises par l'étude ou l'expérience* » (Lacolare, 2011). Une autre définition a été proposée dans (Ermine, 2000) : « *la connaissance est (donc) le résultat de l'interprétation de l'information dans son référentiel. Ce référentiel comprend la sémantique que l'individu attache à l'information et au contexte dans lequel est intégrée cette information* ».

Au niveau industriel, la connaissance est considérée aujourd'hui comme « *une ressource stratégique pour la productivité croissante ; un facteur de stabilité dans un environnement instable et dynamique ; et c'est un avantage concurrentiel décisif* » (Ermine, 2000) pour les individus et les organisations. En effet, l'économie mondiale est en train de passer d'un système de valeur basée sur la production matérielle à un système basé sur les connaissances (Chen, 2010) d'où l'importance croissante accordée au concept d' « entreprise basée sur la connaissance ».

### I.1.2 TYPES DE CONNAISSANCE

Dans la littérature, il existe différentes manières de classifier les connaissances dans les organisations (Figure I.1), en tenant compte par exemple de critères tels que l'expression, l'évidence, l'étendue et la portée de la connaissance (Lai, 2007).

Expression	Evidence	Etendue	Portée
<ul style="list-style-type: none"> <li>• Déclarative</li> <li>• Procédurale</li> </ul>	<ul style="list-style-type: none"> <li>• Tacite</li> <li>• Explicite</li> </ul>	<ul style="list-style-type: none"> <li>• Interne</li> <li>• Externe</li> </ul>	<ul style="list-style-type: none"> <li>• Générale</li> <li>• Spécifique</li> </ul>

Figure I.1. Différentes types de connaissance

#### I.1.2.1 Connaissances déclaratives et connaissances procédurales

Le critère « expression » (Figure I.1) permet de distinguer les connaissances déclaratives (ou connaissances statiques) et les connaissances procédurales (ou connaissances dynamiques). En effet, les connaissances déclaratives décrivent des entités statiques du domaine et les relations potentielles entre ces entités, alors que les connaissances procédurales modélisent les stratégies d'utilisation de cette connaissance statique et décrivent le comportement dynamique des processus et/ou procédures (Ermine et al., 1996).

### **I.1.2.2 Connaissances tacites et connaissances explicites**

Dans le domaine de la gestion des connaissances, une classification classique de la connaissance distingue les connaissances tacites (appelées également « connaissances implicites ») des connaissances explicites (Dienes & Perner, 1999) (Nonaka & Takeuchi, 1995). Cette classification s'exprime dans le critère « évidence » de la connaissance (Figure I.1).

Même si la connaissance est la base de toute activité humaine, seule une partie de celle-ci (i.e. la « connaissance explicite ») est facilement accessible, formalisable, stockable et réutilisable. De ce fait, l'ingénierie des connaissances a récemment été l'objet d'une attention accrue dans les entreprises afin de rendre explicite la connaissance implicite (Lu & Ruan, 2007) (Schreiber, 2008) (Stewart, 1997).

La connaissance implicite concerne les connaissances de l'expérience de soi, i.e. les croyances, les perspectives, les compétences acquises ou les modèles mentaux dans l'esprit des personnes, et est aussi appelée expertise subjective (Nonaka et al., 1996). Ce type de connaissance est souvent emmagasiné dans la tête des employés et est généralement difficile à formaliser et à communiquer. Par contre, la connaissance explicite concerne des connaissances déjà exprimées formellement en utilisant par exemple langages, symboles, règles, objets ou équations, i.e. des données quantifiées, procédures, principes, modèles mathématiques, etc. (Nemati et al., 2002).

### **I.1.2.3 Connaissances internes et connaissances externes**

D'un point de vue industriel, ces deux types de connaissances sont liés à l'« étendue » de la connaissance, i.e. nous parlons de connaissances internes en faisant référence aux connaissances propres de l'entreprise ; en revanche les connaissances externes concernent les connaissances qui ne font pas partie du patrimoine intellectuel de l'entreprise. Elles correspondent au marché, à l'environnement extérieur de l'organisation (Tisseyre, 1999).

### **I.1.2.4 Connaissances générales et connaissances spécifiques**

Ces deux types de connaissances sont liés à la « portée » de la connaissance. Les connaissances spécifiques sont relatives aux connaissances attachées à une situation particulière et dépendent de facteurs divers définissant un contexte, tels que le domaine, le lieu, le temps, l'environnement, etc. Les connaissances générales sont souvent issues de l'analyse de plusieurs connaissances spécifiques et ne dépendent pas d'un contexte particulier (Lai, 2007).

### **I.1.2.5 Synthèse**

Les différents types de connaissances précédemment abordés sont souvent liés entre eux et sont complémentaires dans les entreprises. Dans un premier temps, nous nous intéressons plus

particulièrement dans notre contexte d'étude au potentiel des *connaissances procédurales* décrivant le comportement des processus, des *connaissances explicites* déjà exprimées et structurées (i.e. disponibles), des *connaissances internes* propres à une entreprise et des *connaissances spécifiques* liées à un contexte ou domaine particulier. Ces connaissances feront l'objet d'un traitement particulier lors de notre proposition (Chapitre II) afin de créer et gérer de nouvelles connaissances plus générales pour procurer un avantage concurrentiel aux entreprises.

Nous présentons ci-dessous les modes classiques de création et de transfert de connaissances qui ont donné lieu à la plupart des approches de gestion des connaissances existantes.

### I.1.3 MODES DE CREATION ET DE TRANSFERT DES CONNAISSANCES

Il existe dans la littérature différents modes ou processus de création et de transfert de connaissances. Une classification classique en gestion des connaissances a été proposée dans (Nonaka & Takeuchi, 1995), incluant quatre modes principaux de création et de transfert de connaissances dans les entreprises, tout en distinguant les connaissances implicites, communément appelées savoir-faire, et les connaissances explicites, appelées savoirs.

Nous présentons dans la Figure I.2 les différents processus de création de connaissances et de transfert, suivis par une description de chacun dans un contexte industriel.

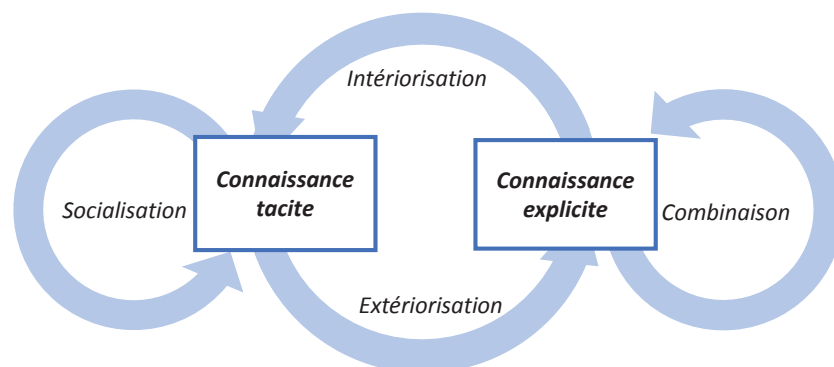


Figure I.2. Processus de création et de transfert de connaissances (Nonaka & Takeuchi, 1995)

- Le processus de **socialisation** (de tacite vers tacite) concerne le partage de la connaissance tacite (souvent dans l'environnement de travail). Il s'agit d'un processus de création de connaissances tacites communes à travers le partage entre acteurs des expériences individuelles acquises au cours des activités antérieures, des compétences techniques ou d'autres formes de connaissances tacites, par l'observation, l'imitation ou la pratique. Autrement dit, il s'agit d'un passage de la connaissance tacite individuelle vers la connaissance tacite du groupe (ou collective).
- Le processus d'**extériorisation** (de tacite vers explicite) considère une articulation de la connaissance tacite vers la connaissance explicite à travers les échanges d'information et de connaissances. Le principal objectif de ce processus dans les entreprises est souvent de formaliser des connaissances tacites à l'aide des experts afin de les rendre explicites, pour une

réutilisation future. Dans ce mode de création ou de transfert de connaissances, nous pouvons considérer par exemple l'acquisition de connaissances dans le développement des « systèmes experts » (Nemati et al., 2002). Dans ce cas, les connaissances sont obtenues, en grande partie, en interviewant les experts du domaine sur la manière avec laquelle ils accomplissent des tâches ou résolvent des problèmes spécifiques (Musen, 1993). Dans cette catégorie, nous pouvons aussi considérer les systèmes qui se concentrent sur la capture et la capitalisation de l'information issue des expériences passées, afin de rendre ces connaissances explicites.

- Le processus de **combinaison** (d'explicite vers explicite) est un processus d'intégration et d'analyse de différents types de connaissances explicites existantes dans une organisation, afin de constituer de nouvelles connaissances, explicites elles aussi. Dans ce contexte, les technologies de l'information jouent un rôle essentiel. Nous pouvons considérer dans ce mode de création de connaissances les techniques statistiques d'analyse de données, ou plus récemment les méthodes d'extraction de connaissances (*data mining*) à partir de l'information existante, que nous présentons plus en détail dans la section I.5. Il peut s'agir en effet de la création de connaissances plus générales à partir de l'analyse des connaissances spécifiques. Des exemples de ce processus sont les connaissances explicites générées à partir des techniques de *data mining* pour supporter la prise de décision, ou encore les connaissances issues du Raisonnement à Partir de Cas (RàPC) pour supporter la résolution de nouveaux problèmes.
- Finalement, le processus d'**intérieurisation** (d'explicite vers tacite) est considéré comme un processus d'apprentissage ou d'appropriation des connaissances explicites diffusées ou partagées dans l'organisation, typiquement avec des supports, documents ou manuels. Il s'agit du passage de la connaissance explicite vers la connaissance tacite, i.e. un processus d'intérieurisation des connaissances explicites par le personnel de l'entreprise.

#### I.1.4 LA GESTION DES CONNAISSANCES AU SEIN DES ORGANISATIONS

La gestion des connaissances est devenue un enjeu très important au sein des organisations, car elle concerne un processus de création de valeur à partir des différents types de connaissances de l'entreprise, qui peuvent être disponibles sous des formes différentes : dans l'esprit des employés, dans les procédures de travail, les normes, les documents internes, les bases de données, les progiciels, etc. (Bergmann, 2002). Le principal objectif des approches de gestion des connaissances est ainsi de modéliser et de stocker les connaissances déclaratives et procédurales d'un domaine donné afin de les réutiliser et de mieux effectuer de nouvelles tâches ou processus.

Dans ce contexte, l'importance de la gestion des connaissances dans les entreprises a induit une diversité d'outils et/ou de démarches pour *repérer*, *préserver*, *valoriser* et *actualiser* les connaissances stratégiques d'une organisation (Grundstein, 2004). Le cycle classique pour la capitalisation des connaissances est présenté dans la Figure I.3. Le processus inclut :

- d'abord, un repérage des sources de connaissances, i.e. il faut les identifier et les localiser ;
- des connaissances cruciales doivent être ensuite préservées ou stockées à l'aide de formalismes de représentation pour faciliter le partage, i.e. il faut les modéliser, les formaliser et les conserver ;



- puis, la valorisation de ces connaissances est liée au fait de pouvoir les réutiliser au sein des organisations, i.e. il faut les rendre accessibles, les diffuser, les exploiter et les intégrer pour créer de nouvelles connaissances ;
- finalement, une mise à jour ou une maintenance est nécessaire afin d'actualiser les connaissances stratégiques de l'organisation, i.e. il faut en permanence les évaluer et les enrichir.

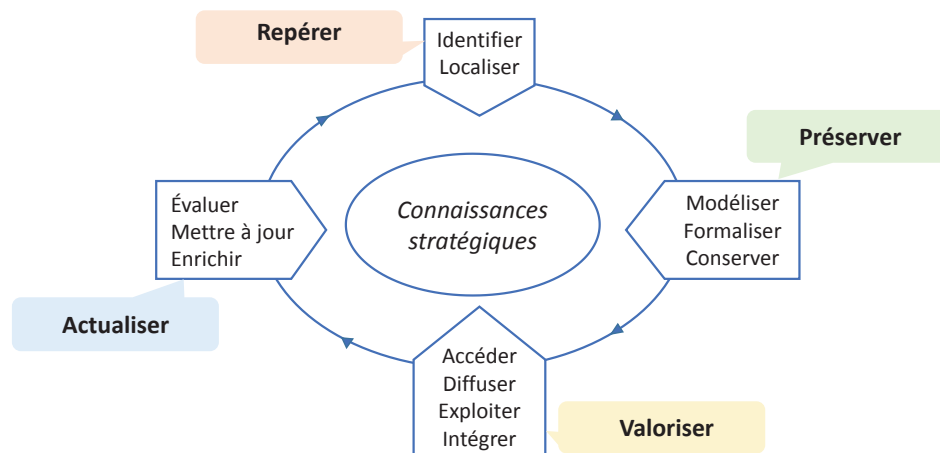


Figure I.3. Cycle classique de la gestion des connaissances selon Grundstein (Grundstein, 2004)

Différentes approches de gestion des connaissances ont été identifiées dans la littérature afin de mettre en œuvre ce cycle de gestion classique. Une classification générale présentée dans (Béler, 2008) (Choo, 1998) (Mom et al., 2007) distingue les méthodes descendantes (*Top-Down*) et les méthodes ascendantes (*Bottom-Up*) de gestion des connaissances, présentées plus en détail ci-dessous.

#### I.1.4.1 Méthodes descendantes

Les méthodes descendantes de gestion des connaissances se concentrent principalement sur les connaissances tacites qui demeurent dans la mémoire des employés de l'organisation, souvent appelés « porteurs » ou « détenteurs » de la connaissance. Un exemple d'approches classiques orientées vers les connaissances tacites concerne, comme nous l'avons déjà signalé, les systèmes experts, pour lesquels l'idée principale est d'extraire des connaissances des experts afin de capturer et de mémoriser leur expertise.

D'autres méthodes d'explicitation des connaissances comme CommonKADS (Common Knowledge Acquisition and Desing Support (Schreiber et al., 1994)), MASK (Method for Analysing and Structuring Knowledge (Matta et al., 2002)), MKSM (Method for Knowledge System Management (Ermine et al., 1996)) ou KOD (Knowledge Oriented Design (Dieng-Kuntz et al., 2001) (Vogel, 1988)) proposent différents modèles de représentation des connaissances et des démarches méthodologiques pour l'acquisition de ces connaissances tacites afin de mettre en place des Systèmes à Base de Connaissances (SBC). Ces approches utilisent des techniques particulières pour extraire des

connaissances expertes (ou tacites) auprès des experts, à l'aide d'un cogniticien, spécialiste de l'ingénierie des connaissances, afin de les formaliser avec des modèles conceptuels et de faciliter le partage et la réutilisation de ces connaissances dans l'organisation.

Cependant, la mise en œuvre de telles méthodologies est un travail lourd, complexe et long à mettre en place puisqu'il demande la mise en œuvre d'une démarche complète de gestion des connaissances (Minor, 2005). En effet, plusieurs difficultés ont été rencontrées pour la mise en place de ces approches ; par exemple, le manque de disponibilité des experts, leur manque de motivation pour partager leur savoir et leur savoir-faire, le besoin d'un cogniticien pour servir d'interface entre les experts et le système, la difficulté des experts à exprimer leurs connaissances, la difficulté du cogniticien à formaliser les connaissances expertes, etc. Du fait de ces difficultés, beaucoup des systèmes experts existants sont relativement simples.

Afin de pallier les difficultés de ces méthodes classiques de gestion des connaissances (ou méthodes descendantes), les méthodes ascendantes constituent une alternative prometteuse.

#### **I.1.4.2 Méthodes ascendantes**

Les méthodes ascendantes visent à construire une base de connaissances à partir des connaissances spécifiques liées aux activités passées d'une entreprise (i.e. fragments de connaissances) en vue de sa réutilisation future par les employés. Cette base de connaissances peut par exemple porter sur les produits, les processus de production, les clients, les stratégies de maintenance, les objectifs stratégiques, etc. Dans ce contexte, d'autres approches ont été aussi largement liées à la gestion des connaissances et aux modes de transfert de connaissances, citons notamment l'apprentissage organisationnel, la mémoire d'entreprise, ou la gestion des expériences que nous aborderons plus en détail dans la section I.2.

La gestion des expériences s'intéresse principalement aux connaissances concrètes issues de chaque résolution de problème et à leur modélisation (Bergmann, 2002). Notamment, en pratique il est plus facile pour les experts de capitaliser les expériences vécues plutôt que des connaissances génériques ayant une portée englobant plusieurs contextes ; un autre avantage de ces approches émergentes basées sur les expériences est lié au fait que leur capitalisation peut être directement intégrée aux processus opérationnels des organisations (Béler, 2008).

## **I.2 GESTION DES EXPERIENCES**

La valorisation des expériences passées est devenue une nécessité stratégique pour les entreprises (Delange & Vogin, 1994) qui commencent à accorder plus d'attention à leurs expériences passées. Leur gestion est souvent définie comme une catégorie spéciale au sein des systèmes de gestion des connaissances. La gestion des expériences est donc considérée comme une méthode de gestion des connaissances ascendante.

Comme la gestion des connaissances, la gestion des expériences (appelée *experience management* en anglais) traite notamment la collecte, la modélisation et le stockage, l'évaluation et l'actualisation de l'expérience (Bergmann, 2002) en vue d'une future réutilisation. En effet, les approches de gestion d'expériences sont surtout dédiées à la gestion des connaissances attachées aux expériences passées (Bergmann, 2002), i.e. des connaissances spécifiques attachées à une activité ou à un contexte particulier. Ces fragments de connaissances concrètes (appelées par la suite expériences) peuvent être analysés a posteriori afin d'extraire de nouvelles connaissances dans les entreprises.

Sur le schéma de la Figure I.4, nous illustrons la notion d'expérience et son potentiel comme source de nouvelles connaissances. D'une part, les expériences passées peuvent faciliter la résolution d'un nouveau problème ; d'autre part, elles peuvent être analysées pour enrichir les connaissances de l'organisation (cette problématique est abordée en détail dans la section I.5). Pourtant, ces fragments de connaissances - ou connaissances spécifiques (les expériences passées) - situées dans un contexte particulier nécessitent un effort pour être généralisées (Jabrouni, 2012).

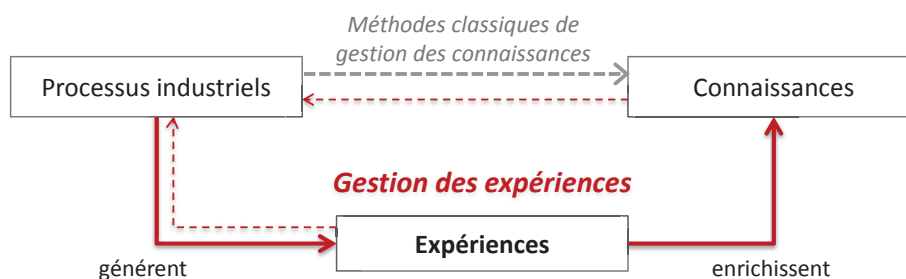


Figure I.4. De l'événement et l'expérience à la connaissance (adaptée de (Rakoto, 2004))

Afin de supporter la prise de décision et d'améliorer les performances des processus industriels, capturer et stocker les expériences passées est souvent considéré comme étant beaucoup plus facile que capturer et stocker des connaissances plus génériques auprès des experts, d'où l'intérêt principal des approches fondées sur les expériences. Toutefois, une connaissance générale a une plus grande portée qu'une expérience spécifique, d'où l'intérêt d'analyser les informations issues de l'expérience (Bergmann, 2002). De plus, pour les acteurs opérationnels, il est souvent plus facile de valider une expertise extraite d'expériences vécues que de décrire une connaissance générique hors contexte (Kolb, 2000).

### I.2.1 DEFINITION D'UNE EXPERIENCE

Dans la littérature, il existe plusieurs définitions d'une expérience. Dans (Apedome et al., 2011), une expérience est définie comme un « *fragment de connaissance lors de processus de résolution de problème* ». D'un point de vue plus général, une autre définition proposée par (Bergmann, 2002) définit une expérience comme une connaissance spécifique située dans un contexte particulier ou dans la résolution d'un problème concret. Nous adoptons dans nos travaux cette dernière définition qui est moins restrictive, en considérant d'autres contextes que celui de la résolution d'un problème, qui est néanmoins le plus couramment rencontré.

Finalement, une relation particulière existe entre connaissance et expérience : une expérience est souvent considérée comme une spécialisation d'instance de la connaissance, ou comme une instance singulière (ou forme) de connaissance antérieure (Sun, 2004).

## **I.2.2 APPROCHES DE GESTION DES EXPERIENCES**

Dans les entreprises, différentes expressions sont utilisées pour décrire des démarches liées aux approches de gestion des expériences. En Europe par exemple, l'expression la plus courante est « *experience management* » (gestion des expériences) (Bergmann, 2002) ; aux Etats-Unis, on parle plutôt de « *lesson learned* » (leçons acquises) et/ou d' « *experience feedback* » (retour d'expérience) (Béler, 2008) (Kamsu-Foguem et al., 2008) (Lejarraga & Gonzalez, 2011). Cette dernière terminologie est aussi la plus utilisée en France et c'est l'expression que nous allons retenir dans ce mémoire, avec comme principaux objets d'intérêt l'expérience et sa gestion.

Dans ce cadre général, nous considérons les méthodes de retour d'expérience comme une catégorie spécifique de la gestion des connaissances, plus particulièrement dédiée à la gestion de la connaissance attachée à l'expérience acquise au cours des activités antérieures.

## **I.2.3 LE RETOUR D'EXPERIENCE**

Le Retour d'Expérience (REx) est une initiative prometteuse visant à éviter la reproduction des erreurs du passé et à réutiliser des solutions éprouvées (Renaud et al., 2008). Cette démarche organisationnelle est devenue un des vecteurs les plus importants de la gestion des connaissances des entreprises (Armaghan, 2009), de la conception de produits à la fabrication, en passant par la maintenance des équipements industriels (Karray et al., 2014).

Dans la littérature, il existe de nombreuses définitions du retour d'expérience et diverses propositions de méthodologies qui varient notamment selon le domaine considéré. Par exemple, dans le domaine de la gestion de risques, le groupement REXAO (Retour d'Expérience et Apprentissage Organisationnel) met l'accent sur le traitement d'événements négatifs (i.e. anomalies et dysfonctionnements) pour en tirer des leçons (nouvelles consignes, modifications organisationnelles, etc.) (Rexao, 2003). Une autre définition elle aussi liée aux événements négatifs et à un contexte de résolution de problèmes a été proposée dans (Bergmann, 2002). La gestion d'expérience est considérée comme « *un type de gestion des connaissances qui se restreint aux connaissances issues de l'expérience, c'est-à-dire, aux connaissances spécifiques relatives à un contexte particulier de résolution d'un problème* ». Aux Etats-Unis, le retour d'expérience s'intéresse plutôt au concept de leçon acquise et considère les événements négatifs mais aussi les événements positifs ; ainsi dans (Bickford, 2000) « *une leçon acquise est une bonne pratique ou une approche innovante qui est capturée et partagée afin d'encourager sa réutilisation. Une leçon acquise peut également représenter un mauvais fonctionnement ou une expérience qui est capturée pour en éviter la récurrence* ».

A partir de l'ensemble des définitions existantes, nous avons retenu la définition suivante qui reprend des aspects essentiels évoqués dans les propositions précédentes : le retour d'expérience

(REx) est une démarche structurée permettant la capitalisation, le traitement et l'exploitation des informations issues de l'analyse d'événements positifs et/ou négatifs (Béler, 2008) (Clermont et al., 2007) (Rakoto et al., 2002), ayant comme point clé le fait de vouloir tenir compte des expériences passées pour en tirer des leçons (Chebel-Morello, 2008). Cette définition considère le traitement des informations issues d'événements positifs et/ou négatifs, présente la démarche comme un processus et fait apparaître un aspect important auquel nous nous sommes fortement intéressés : la notion d'expérience comme une source de nouvelles connaissances.

Une démarche de REx est un projet ambitieux à mettre en place au sein des organisations car elle met en œuvre un ensemble de ressources diverses, notamment humaines et technologiques, qui doivent être bien gérées afin de réduire les répétitions d'erreurs et de favoriser les bonnes pratiques dans l'organisation (Rakoto, 2004).

#### **I.2.4 TYPES DE RETOUR D'EXPERIENCE**

Afin de positionner nos travaux dans le contexte du retour d'expérience, nous considérons deux manières principales de classer le REx dans les entreprises : une qui prend en compte la nature des événements traités, et l'autre qui considère la façon de traiter ces événements passés (Villeneuve, 2012).

##### **I.2.4.1 REx positif et REx négatif**

La distinction entre REx positif et REx négatif repose sur la nature des événements considérés. Le REx qualifié de positif favorise dans la démarche les événements « positifs » ou les « bonnes pratiques ». A l'opposé, le REx négatif ne considère que les événements « négatifs » (i.e. accidents, incidents, anomalies, défaillances, problèmes subis, ...) ou toute situation considérée comme « anormale ». Ce dernier type de REx est le plus considéré dans la littérature, en particulier par les différentes méthodes de résolution de problèmes, qui peuvent en effet faciliter la capitalisation des expériences passées.

##### **I.2.4.2 REx statistique et REx cognitif**

D'autre part, nous pouvons distinguer le REx statistique et le REx cognitif selon la manière avec laquelle le système traite les informations issues des événements passés et selon la quantité de données manipulées. Le premier utilise des techniques statistiques sur les informations capitalisées, surtout lorsqu'il s'agit de quantités d'informations importantes. Ce cas correspond à la réalité actuelle du monde industriel, de par la généralisation des systèmes d'information comme les ERP<sup>1</sup> (*Enterprise Resource Planning*) (Botta-Genoulaz et al., 2005) qui permettent d'accumuler des

---

<sup>1</sup> ERP a été traduit en français par l'acronyme PGI (Progiciel de Gestion Intégré).

quantités importantes de données. Ce type de retour d'expérience est souvent considéré comme un retour d'expérience « élémentaire », utilisant des méthodes mathématiques pour analyser les événements passés. Néanmoins, des améliorations considérables ont eu lieu dans ce domaine avec les avancées des technologies de l'information et des techniques d'analyse de données plus performantes, visant à extraire de la connaissance à partir de l'information disponible (section I.5).

Contrairement au REx statistique, le REx cognitif se focalise sur les analyses des événements par les experts afin de mettre en évidence les points importants des expériences passées, malgré l'absence de données statistiques (Rakoto, 2004) (Villeneuve, 2012).

### I.3 MISE EN ŒUVRE D'UN SYSTEME DE RETOUR D'EXPERIENCE

Nous présentons sur la Figure I.5 notre schéma du processus de REx inspiré des travaux de thèse de Rakoto (Rakoto, 2004), qui est cohérent avec les étapes classiques de gestion des expériences proposées dans (Bergmann, 2002) :

- capturer, collecter et stocker les informations qui caractérisent les expériences passées (la **phase de capitalisation du REx**) ;
- évaluer, adapter, transformer les expériences en nouvelles connaissances (la **phase de traitement du REx**) ;
- réutiliser les connaissances et maintenir la base (la **phase d'exploitation du REx**).

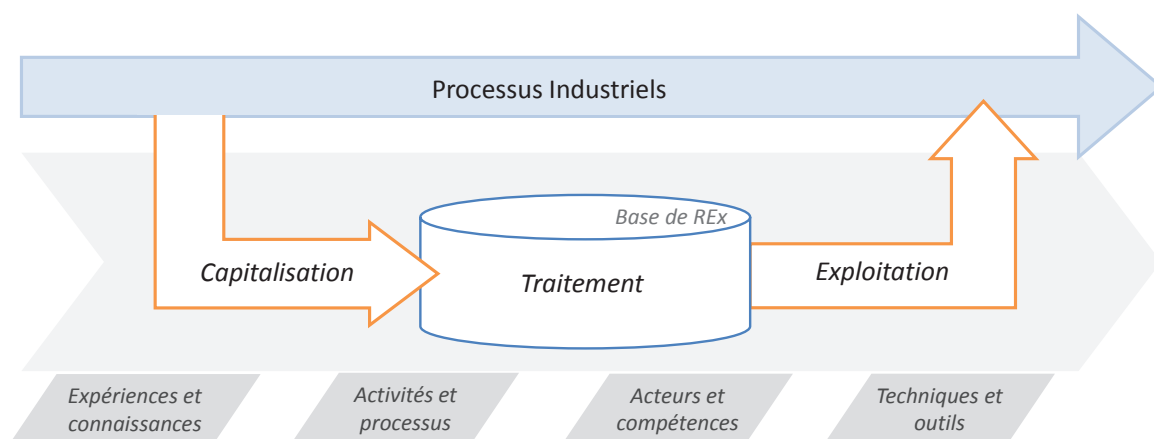


Figure I.5. Structuration des processus et des composants supports du REx

Cette démarche générale place l'existence d'une base d'expériences, appelée souvent base de retour d'expérience ou base de REx, au cœur du processus. Elle contient à la fois des expériences et des connaissances plus génériques (Béler, 2008).

Afin de structurer cette démarche dans un domaine industriel, quatre dimensions essentielles sont prises en compte tout au long du processus (Figure I.5) : notamment les *expériences* et les *connaissances* manipulées, les *activités* et les *processus* internes de l'organisation, les *acteurs* qui mettent en application leurs analyses et *compétences* lors de chacune des étapes, et les *techniques* et *outils* mis en œuvre pour instrumenter ou opérationnaliser le processus global. Ces dimensions ou composants support sont abordés plus en détail ci-dessous.

### **I.3.1 DIMENSION « EXPERIENCES ET CONNAISSANCES »**

La démarche de retour d'expérience manipule des expériences mais aussi d'autres types de connaissances. Nous allons rencontrer ainsi autant des expériences que des connaissances dans la base de REx, qui nécessitent d'être formalisées afin de faciliter leur traitement et leur future réutilisation. Cependant, cette formalisation dépend essentiellement du système et de sa complexité, aussi bien que des futurs utilisateurs (experts du domaine d'application). En effet, la définition d'un formalisme pour représenter les connaissances dans la base est considérée comme un aspect clé dans la gestion des expériences/connaissances. Nous allons présenter dans la partie suivante les principaux formalismes de représentation des connaissances susceptibles d'être utilisés dans ce contexte (section I.4).

### **I.3.2 DIMENSION « ACTIVITES ET PROCESSUS »**

La démarche globale de retour d'expérience considère trois processus principaux (la capitalisation, le traitement et l'exploitation), qui comportent plusieurs sous-processus ou activités selon le type d'application réalisée et selon les approches/méthodes complémentaires utilisées.

### **I.3.3 DIMENSION « ACTEURS ET COMPETENCES »**

Tout au long du processus global de REx, l'implication de différents acteurs est fondamentale, de la conception du système jusqu'à son exploitation et à sa maintenance (i.e. opérateurs, experts du domaine d'application, cognitiens, administrateurs, etc.). Ces acteurs humains jouent un rôle fondamental dans le processus, d'où l'importance de définir leurs rôles et leurs responsabilités lors de l'utilisation du système.

### **I.3.4 DIMENSION « TECHNIQUES ET OUTILS »**

Pour la mise en œuvre et l'opérationnalisation de la démarche générale, des techniques et des outils d'implémentation sont aussi nécessaires. Afin de mieux gérer et traiter les informations qui caractérisent une expérience, il est en effet nécessaire de les modéliser pour permettre l'exploitation de ces fragments de connaissances.

Des techniques classiques comme le Raisonnement à Partir de Cas (RàPC) (Aamodt & Plaza, 1994), ont fait l'objet d'un intérêt croissant au cours des dix dernières années. Cette technique, ayant ses origines dans le raisonnement par analogie, permet d'adapter la solution d'un problème déjà résolu pour résoudre un nouveau problème (Lieber, 2007) à partir d'une base d'expériences, appelée base de cas dans le RàPC. Il s'agit d'une technique attrayante dans la démarche de retour d'expérience lors de la réutilisation des expériences contenues dans la base de REx. Bien que cette technique permette d'adapter des fragments de connaissances, elle ne permet pas de généraliser ces connaissances spécifiques stockées dans la base, ce qui peut parfois être fait par des techniques avancées d'analyse de données. Nous proposons ainsi un retour d'expérience plus complet intégrant des techniques émergentes pour générer des connaissances plus généralisées à partir des informations issues de l'expérience (Chapitre II).

Nous pouvons aussi considérer dans cette section les outils qui assurent l'interaction entre l'utilisateur et le système, i.e. les parties visibles de l'application (interfaces) liées notamment à l'exploitation des expériences et des connaissances contenues dans la base de REx.

## I.4 FORMALISATION DES CONNAISSANCES

La formalisation des connaissances est un mécanisme indispensable pour valoriser et faciliter la réutilisation future des connaissances capitalisées. En effet, il s'agit d'un moyen de partager les connaissances en favorisant leur compréhension au moyen de représentations explicites (Rakoto, 2004) et par un traitement plus automatisé de l'information.

Cette partie est consacrée aux principaux formalismes de représentation de connaissances susceptibles d'être utilisés dans notre travail pour la représentation des différents types de connaissances liés aux expériences passées. Le choix d'un formalisme spécifique sera justifié par le fait que notre approche est destinée à un processus de recherche et d'exploitation des expériences passées ainsi qu'à un traitement d'extraction, d'évaluation et de validation de nouvelles connaissances.

Avant d'aborder les formalismes possibles pour la représentation des connaissances, il nous semble important d'exposer quelques points clés permettant de prendre en compte les différents types de connaissances disponibles et les besoins des entreprises (Lai, 2007) :

- *Une représentation unifiée* : les entités et les relations utilisées pour la représentation de la connaissance déclarative doivent être aussi utilisées pour formaliser la connaissance procédurale.
- *Une représentation suffisamment intuitive et naturelle* : le formalisme de représentation de connaissances devrait faciliter le rôle des experts du domaine industriel autant pour exprimer que pour comprendre la connaissance, car ces experts ne sont pas a priori informaticiens, cognitivistes ou spécialistes de l'ingénierie des connaissances.
- *Un vocabulaire commun* : l'interprétation d'une entité ou concept devrait être la même pour les différents utilisateurs.



Nous trouvons dans la littérature de nombreux formalismes de représentation des connaissances utilisés dans les systèmes de gestion des expériences/connaissances. Ci-dessous, nous présentons une synthèse de ces formalismes en distinguant les approches basées sur les *frames* (Borgida, 1996) (i.e. les Frames et les Logiques de Description) et celles basées sur les *graphes* (Sowa, 1984) (i.e. les Réseaux Sémantiques et les Graphes Conceptuels).

#### I.4.1 APPROCHES BASEES SUR LES FRAMES

- Les **Frames** (Minsky, 1975) sont un premier formalisme informatique dans lequel les concepts représentent des ensembles d'objets ayant des propriétés communes. Un *frame* (schéma ou cadre) est donc une structure représentant une situation typique comportant plusieurs attributs. Minsky (Minsky, 1975) a proposé de structurer la connaissance à travers un ensemble de *frames*, différents types d'informations étant liés à chaque frame, les attributs étant représentés par un ensemble de facettes. La principale critique à ce formalisme est liée à une expressivité limitée et au fait que, selon les logiciens, il ne permet pas d'inférer des connaissances valides (Apedome et al., 2011).
- Les **Logiques de Description** (LDs) (Borgida, 1996) sont issues de la théorie des *frames* et sont couplées avec les principes des réseaux sémantiques que nous exposons ci-dessous. Ce formalisme peut être utilisé pour représenter la connaissance d'un domaine d'application d'une manière structurée et formelle. En effet, on peut distinguer dans la plupart des langages de ce type deux niveaux de représentation des connaissances : le niveau terminologique (qui définit les concepts ou connaissances générales) et le niveau factuel (ou assertionnel, qui définit des connaissances spécifiques). Ces niveaux ont donné naissance aux notions de *T-Box* et *A-Box* que l'on retrouve dans la plupart des logiques de description. Il s'agit de l'une des plus importantes familles de formalismes de représentation de connaissances basés sur la logique de premier ordre. Malgré cela, ce formalisme a perdu certaines des qualités graphiques et/ou visuelles de ses ancêtres (i.e. les *frames* et les réseaux sémantiques) (Dau & Eklund, 2008).

#### I.4.2 APPROCHES BASEES SUR LES GRAPHS

- Les **Réseaux Sémantiques** (Quillian, 1968) ont été proposés en 1961 puis en 1968 par Quillian afin de construire un modèle de la « mémoire humaine » fondé sur un réseau sémantique de mots construit à partir d'expériences en psycholinguistique. Cette approche introduit l'utilisation de nœuds pour représenter les entités conceptuelles. Un réseau sémantique est donc un graphe orienté et étiqueté, constitué d'un ensemble de nœuds et d'arcs, qui permet la représentation des relations sémantiques entre les entités conceptuelles.

Les nœuds représentent les entités conceptuelles qui sont interconnectées par des arcs représentant des liens ou des relations conceptuelles. Ce formalisme a aussi introduit l'idée d'organiser les concepts dans des hiérarchies en considérant que certains concepts sont généraux et d'autres plus spécifiques.

La principale critique à ce formalisme, très présente dans la littérature, est le manque de sémantique formelle clairement définie (Chein & Mugnier, 2008), point gênant pour toute interprétation. En conséquence, ce formalisme a donné lieu à de nombreux modèles de représentation des connaissances dotés d'une sémantique formelle, parmi lesquels les graphes conceptuels (Chein & Mugnier, 1992).

- Les **Graphes Conceptuels** (GCs) (Sowa, 1984) et les Logiques de Description (LDs) sont tous deux des descendants des réseaux sémantiques et sont basés sur la logique des prédicats du premier ordre. Toutefois, une représentation de la connaissance à l'aide des GCs offre des facilités spécifiques d'explicitation des connaissances.

Les GCs sont des graphes finis, connectés et bipartis (Sowa, 1984), incluant deux types de nœuds : les concepts et les relations conceptuelles. Ces deux types de nœuds définissent une structure ontologique permettant de construire les différents graphes conceptuels nécessaires à la représentation de la connaissance. Ce formalisme permet donc une séparation nette entre les connaissances ontologiques (structure ontologique appelée « support », formée d'une hiérarchie de concepts et d'une hiérarchie de relations), et les connaissances factuelles représentées par les graphes (Baget, 2001), i.e. les connaissances déclaratives et procédurales. Pour l'inférence, le mécanisme privilégié des GCs est l'opération de projection (Sowa, 1984).

Ce formalisme et ses mécanismes sont intéressants dans le contexte du REx, car ils expriment les connaissances sous une forme graphique, logique, précise, compréhensible par des utilisateurs et adaptée à un traitement automatique.

Dans la suite de ce mémoire, nous nous focalisons sur l'étude d'une problématique liée à la démarche de REx permettant d'enrichir sa mémoire. Nous allons étendre la démarche classique de retour d'expérience en y ajoutant un processus d'Extraction de Connaissances à partir de Données (ECD).

## I.5 L'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES

Les progrès des technologies de l'information et de la technologie de stockage ont poussé les travaux de recherche vers la recherche de connaissances à partir des bases de données disponibles dans les entreprises. Une implémentation croissante des systèmes d'information en milieu industriel, tels que les ERP, rend en effet disponible une quantité d'information toujours plus grande, car aujourd'hui même des opérations simples sont enregistrées dans le système d'information. Cependant, il a été constaté que ces bases de données sont en général peu ou pas exploitées au sein des organisations (Bendoly, 2003) (Marban et al., 2009) (Kamsu-Foguem & Mathieu, 2014).

Des méthodes assez performantes concernant le traitement statistique des données pourraient être appliquées sur les informations recueillies ; néanmoins, ces méthodes ont montré leurs limites quant à l'interprétation des résultats obtenus (Renaud et al., 2008). Des méthodes issues de l'intelligence artificielle relatives à la fouille de données ont aussi été développées afin d'extraire de nouvelles connaissances en utilisant des techniques avancées d'analyse de données (Fayyad et al., 1996). Il

s'agit de méthodes prometteuses pour les organisations, qui s'intéressent de plus en plus à l'exploitation de leurs bases de données afin d'extraire des motifs ou modèles intéressants et utiles. Dans ce contexte, il existe actuellement des applications industrielles utilisées dans la communauté d'ECD qui proposent un guide de mise en œuvre d'applications de *data mining* ; citons notamment CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Chapman et al., 2000) et SEMMA (*Sample, Explore, Modify, Model and Assess*) (SAS Institute Inc, 2013).

### I.5.1 DEFINITION DU PROCESSUS D'ECD

Au-delà des méthodes d'analyse de données traditionnelles, les approches d'extraction de connaissances ont développé depuis quelques années de nouveaux outils « intelligents » pour améliorer la découverte de nouvelles connaissances dans différents domaines industriels.

L'Extraction de Connaissances à partir de Données (ECD<sup>2</sup>), plus communément appelée *data mining*, est un processus non trivial d'identification de motifs ou modèles *valides, nouveaux*, potentiellement *utiles* et *compréhensibles* à partir des données (Fayyad et al., 1996). Il s'agit d'un processus interdisciplinaire qui combine différentes approches issues notamment de l'intelligence artificielle, de l'informatique, de l'apprentissage automatique, de la gestion de bases de données, de la visualisation de données et/ou des algorithmes mathématiques et statistiques (Liao, 2003) afin de supporter les processus de prise de décision dans les entreprises.

Etant donné le potentiel important des données recueillies en tant que sources de nouvelles connaissances (Harding et al., 2005), l'ECD est devenue essentielle dans de nombreux domaines industriels, incluant la conception de produits et de processus, la planification des ressources, le contrôle qualité, la maintenance, la gestion de la relation client, les systèmes d'aide à la décision, etc., car susceptible de procurer des avantages concurrentiels importants (Zighed & Rakotomalala, 2002).

Dans ce contexte, nous devons considérer les composants basiques pour le développement d'un tel système d'extraction de connaissances. Les techniques de *data mining* sont au cœur de ces systèmes, puisqu'elles permettent l'extraction de nouvelles connaissances ; les entrées du système sont généralement les informations issues des bases de données, les sorties étant les connaissances extraites qui devraient être interprétées et évaluées.

Les masses d'informations stockées dans les bases de données décrivent un grand nombre de transactions qui contiennent de nombreux attributs. Dans une base de données relationnelle par exemple, les données sont généralement organisées dans des tables de transactions, chaque transaction contenant une liste ordonnée de valeurs (une valeur pour chaque attribut). La plupart des techniques d'ECD visent à l'exploitation simultanée des valeurs d'attributs disponibles afin de découvrir de nouvelles connaissances.

Des exemples classiques de types de connaissances extraites sont les relations issues de l'analyse des données pour détecter les défaillances dans les équipements (Baohui et al., 2011) dans le

---

<sup>2</sup> Le terme originel en anglais est « *Knowledge Discovery in Databases* » (KDD).

domaine de la maintenance, ou les relations permettant de trouver les machines responsables des défauts de qualité dans les produits (Chen et al., 2005) dans le domaine de la qualité.

## **I.5.2 PRINCIPALES ETAPES D'UN PROCESSUS D'ECD**

Parmi les différents étapes du processus d'ECD, le *data mining* est considéré comme une étape essentielle, consistant à appliquer des algorithmes d'analyse de données afin de générer de nouvelles connaissances (Köksal et al., 2011). Il s'agit en effet d'une technique offrant de nouvelles théories et méthodes pour la découverte de connaissances cachées (relations, motifs ou modèles) à partir d'importants volumes d'informations (Harding et al., 2005).

Le processus classique d'ECD comporte principalement trois étapes : la préparation des données (*prétraitement*), l'application des algorithmes d'extraction (*data mining*), et finalement l'évaluation des résultats obtenus (*post-traitement*).

### **I.5.2.1 Prétraitement**

Cette étape est indispensable dans le processus d'ECD et requiert une attention importante afin de disposer de données fiables avant l'application des algorithmes d'extraction, garantissant ainsi dans une certaine mesure la qualité des résultats.

On peut rencontrer différentes difficultés dans l'exploitation des bases de données disponibles pour l'extraction de connaissances. Ces difficultés proviennent du fait que les bases de données du monde réel sont généralement dynamiques, incomplètes, bruitées et contiennent de plus en plus d'informations (Frawley et al., 1992). D'autres préoccupations sont relatives à la question de savoir si la base de données contient des informations adéquates et pertinentes pour l'exploitation. Pour cela, considérer les techniques de *nettoyage*, de *discrétisation*, de *transformation* ou de *réduction* des données avant l'étape de *data mining* devient essentiel afin d'améliorer la qualité des données disponibles, puis des connaissances générées. Il s'agit d'une manière générale de comprendre les propriétés des données, de supprimer les données peu intéressantes, d'enrichir les données par des compléments d'informations, de fractionner ou créer de nouveaux attributs, et/ou d'en combiner d'autres.

Nous présentons ci-dessous certains techniques de prétraitement des données susceptibles d'être utilisées en fonction des données disponibles et des formats des données d'entrée demandés par les algorithmes de *data mining*.

#### **Nettoyage des données**

Le nettoyage des données (*data cleaning* en anglais) consiste à détecter, corriger et éliminer les erreurs, les inexactitudes et/ou les incohérences des données (Rahm & Do, 2000). Certaines incohérences dans les données pourraient être corrigées manuellement, cependant, quand il s'agit de grandes quantités de données, il devient nécessaire d'utiliser des systèmes automatiques ou semi-automatiques afin de détecter des difficultés et de corriger certains problèmes comme les

données manquantes ou aberrantes, les erreurs lors de saisie des informations par les opérateurs ou les données non valides. Le principal objectif de cette étape est donc de générer des bases de données modifiées (soit en inférant des données soit en en corrigeant d'autres) avant l'application des algorithmes d'analyse.

Il est nécessaire de définir des règles pour gérer ou pour corriger les problèmes présents dans les données. Plusieurs solutions ont été proposées. Par exemple dans le cas des données numériques, des données manquantes peuvent être remplacées par la valeur la plus fréquente de l'attribut concerné, ou elles peuvent être estimées par des méthodes d'induction (Zighed & Rakotomalala, 2002). D'autre part, pour le traitement des données aberrantes, des règles ou stratégies de traitement doivent être préétablies afin de détecter les données qui peuvent gêner l'analyse. Des solutions ont été aussi proposées dans le cas des données qualitatives (Hellerstein, 2008) ; toutefois, la plupart des méthodes de nettoyage se concentrent surtout sur les erreurs dans les attributs de type quantitatif.

### Intégration des données

Un autre contexte qui demande un prétraitement de données provient du fait d'avoir plusieurs sources de données, i.e. des données localisées sur des sites différents (bases de données relationnelles, entrepôts de données, sources externes, etc.) (Han & Kamber, 2006). Dans ce contexte, l'objectif est d'intégrer les bases de données concernées dans une nouvelle base qui regroupe toute l'information pertinente pour faciliter l'application des techniques de *data mining* (Figure I.6).

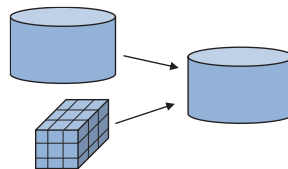


Figure I.6. Intégration de données avant l'application de techniques de data mining

### Réduction des données

En outre, il y a généralement dans les bases de données des attributs ou même des transactions qui ne sont pas intéressantes ou significatives dans un contexte donné. De ce fait, il est possible de réduire le volume de données afin de ne considérer que celles qui sont les plus utiles et intéressantes à analyser en fonction de l'objectif du processus d'extraction. On peut ainsi optimiser la mémoire et le temps d'analyse des algorithmes tout en maintenant l'intégrité des données d'origine (Figure I.7). Dans ce but, des méthodes statistiques d'échantillonnage, de sélection de transactions ou de sélection d'attributs peuvent être considérées selon des conditions exprimées par l'utilisateur (Zighed & Rakotomalala, 2002).

Parmi les différentes techniques de réduction de données, la discrétisation est une forme de réduction consistant à transformer des attributs (en les séparant), ou à réduire le nombre de niveaux

d'un attribut (en remplaçant les concepts de bas niveau par des concepts de haut niveau d'une hiérarchie de concepts).

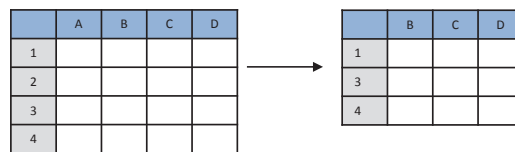


Figure I.7. Réduction de données avant l'application de techniques de data mining

### Transformation des données

Généralement, les données issues des bases de données ne sont pas exploitables directement par les techniques de *data mining*, c'est pourquoi une transformation des données initiales est en général nécessaire afin de garantir que celles-ci soient dans le format d'entrée demandé par l'algorithme de *data mining* utilisé (Figure I.8). Tout dépend donc de la technique choisie, certaines étant plus contraignantes que d'autres.

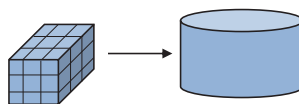


Figure I.8. Transformation des données avant l'application de techniques de data mining

#### I.5.2.2 Data Mining

Le *data mining* est une étape fondamentale du processus d'extraction de connaissances. Les systèmes de *data mining* sont classifiés suivant différents critères tels que le type de bases de données exploitées, le type de connaissances extraites à partir des données, le type de technique utilisée ou le domaine d'application (Han & Kamber, 2001). Dans le contexte industriel, ces systèmes sont souvent classés en deux groupes principaux (Han & Kamber, 2006) en fonction de la nature des données considérées ainsi que du type de connaissances extraites : *data mining descriptif* et *data mining prédictif*.

Le *data mining descriptif* (Dasu & Johnson, 2003) (Giudici, 2003) se concentre principalement sur la découverte de modèles ou de relations afin de décrire les données. Le *data mining prédictif* se concentre surtout sur la prédiction du comportement d'un modèle, et détermine les valeurs futures des variables à partir de l'information issue des bases de données (Choudhary et al., 2009).

Aujourd'hui, une grande variété de techniques de *data mining* existe ; nous nous concentrons donc ici sur les plus importantes. Un aperçu général des techniques est d'abord présenté, puis est détaillé suivant le choix de la technique et des algorithmes à utiliser dans notre démarche étendue de REX, en tenant compte des objectifs du processus d'extraction, des données disponibles à analyser et des attentes de l'utilisateur du domaine.

- La **classification** est une technique de type prédictif qui affecte les objets d'une base de données à une des classes prédéfinies (Fayyad et al., 1996) selon un modèle (ou fonction) de classification préétabli. Le principal objectif est donc ici d'utiliser le modèle (règles de classification, arbres de décision, etc.) pour prédire la classe de chaque objet et/ou concept analysé (Han & Kamber, 2006) dans la base de données, i.e. chaque objet va être associé à un nom de classe.

De nombreuses applications relèvent de cette problématique, comme le diagnostic médical, la gestion des stocks ou le ciblage de clientèle (Ngai et al., 2009). Un exemple classique de cette technique concerne l'étude des marchés financiers, afin de classifier de manière automatique les futurs demandeurs de prêts et ainsi aider la prise de décision dans l'entreprise.

- Le « **clustering** », appelé également regroupement ou classification non supervisée, cherche à identifier et à créer des sous-ensembles homogènes d'objets, catégories ou groupes (i.e. des *clusters*) pour décrire les données, à partir d'un ensemble total hétérogène d'objets dans la base de données. Contrairement à la technique de classification présentée précédemment, le *clustering* analyse les données sans connaître a priori les classes ou groupes. Cette technique de type descriptif permet en effet de fractionner la base de données en groupes d'objets ayant des caractéristiques similaires, e.g. en utilisant le principe de la similarité, qui est généralement mesurée par des fonctions de distance entre les objets. La technique a pour objectif de maximiser la similarité intra-groupe tout en minimisant la similarité entre groupes différents (Han & Kamber, 2001).

Ces techniques sont souvent classées en plusieurs catégories : les méthodes de partitionnement, les méthodes hiérarchiques, les méthodes basées sur la densité, les méthodes basées sur la grille, et les méthodes fondées sur un modèle (Choudhary et al., 2009). Un exemple de *clustering* dans le domaine du marketing inclut la découverte de différentes catégories homogènes pour représenter des groupes cibles de consommateurs (Huang et al., 2007).

- La « **summarization** », appelée également caractérisation, considère l'identification et la présentation des caractéristiques générales ou propriétés typiques d'un sous-ensemble de données, en essayant de trouver une description concise pour chaque sous-ensemble (Fayyad et al., 1996) (Lee & Kim, 1997). Un exemple simple de *summarization* concerne le calcul de la moyenne et des écarts-types pour un ensemble de données, souvent utile pour l'analyse ou la visualisation des données.
- Les **règles d'association** concernent une technique de *data mining* de type descriptif qui cherche à identifier des groupes d'*items* (attributs et leur valeur) qui se produisent ensemble. Le problème d'extraction de règles d'association a été introduit par Agrawal (Agrawal et al., 1993) dans le domaine des transactions de ventes, chacune constituée d'une liste d'articles achetés, afin de découvrir des relations entre les articles le plus fréquemment achetés ensemble. Ces relations ne sont pas basées sur des propriétés inhérentes aux données (comme les dépendances fonctionnelles), elles sont plutôt basées sur la co-occurrence des *items* dans les données de la base (Choudhary et al., 2009). Une règle dans ce domaine sera par exemple : « les clients qui achètent du lait ont une tendance à acheter des céréales ».

Nous allons aborder plus en détail cette problématique d'extraction de règles d'association (type de connaissances extraites sur lequel porte notre processus d'ECD) et ses applications dans la section I.5.3.



- La **régression** est une technique de type prédictif qui associe un élément des données à une variable de prédiction. Dans ce contexte, la régression linéaire concerne la méthodologie statistique la plus utilisée pour la prévision numérique (Han & Kamber, 2006). En revanche, la régression logistique est la principale méthode de prédiction pour les variables qualitatives (Giudici, 2003). D'autres techniques associées à la prédiction sont entre autres les arbres de décision, les réseaux bayésiens ou les réseaux de neurones.
- Finalement, la **découverte de motifs séquentiels** est une technique de type descriptif d'identification d'associations ou de modèles séquentiels au fil du temps (Agrawal & Srikant, 1995). Il s'agit de la détection de motifs séquentiels fréquents ou périodiques à partir d'une base de séquences, dans laquelle chaque séquence est décrite par une liste ordonnée d'ensembles d'*items*. Un exemple classique de cette méthode concerne l'identification de la séquence d'achat d'un client afin de prédire la suite de cette séquence.

### I.5.2.3 Post-traitement

Le post-traitement, appelé également *post-mining*, concerne la dernière étape du processus d'ECD. En effet, les modèles ou motifs extraits ne devraient pas être utilisés directement en toute confiance, il faut d'abord une étape d'analyse et d'interprétation de l'information découverte (Giudici, 2003) afin d'estimer sa qualité. Il est essentiel que les modèles soient évalués en termes de qualité et/ou d'efficacité avant leur utilisation possible sur le terrain (Larose, 2005).

Plus particulièrement, cette dernière étape vise l'évaluation, la visualisation et la validation par l'utilisateur des résultats issus de l'étape du *data mining*, afin de pouvoir garantir les résultats extraits comme de nouvelles « connaissances » dans le domaine.

Dans de nombreux cas, il est possible de définir des mesures statistiques, objectives ou d'utilité pour définir l'intérêt d'un modèle extrait. D'autres notions telles que leur nouveauté et leur compréhensibilité sont de plus en plus prises en compte dans cette étape, mais concernant des aspects beaucoup plus subjectifs qu'on abordera plus en détails dans le Chapitre III, dédié à cette étape de post-traitement dans le processus d'ECD.

## I.5.3 EXTRACTION DE REGLES D'ASSOCIATION

Dans cette section, nous présentons les généralités et les principales motivations de la technique de *data mining* choisie pour le développement d'un processus d'extraction de connaissances lié à la démarche de REx.

### I.5.3.1 Motivations

Comprendre les relations entre les paramètres, les variables ou les attributs d'un système ou d'un processus a été depuis toujours un défi majeur dans les entreprises. Ce problème a été initialement résolu en utilisant des modèles mathématiques qui prennent en compte les hypothèses initiales.



Cependant, la complexité croissante de tels systèmes industriels et le progrès des technologies de l'information rendent disponibles aujourd'hui un grand volume d'information qui est souvent inexploitable par les techniques d'analyse de données « traditionnelles » en raison de sa dimension.

Comme nous l'avons évoqué, l'idée d'extraire relations, corrélations ou associations à partir des données a commencé notamment dans le contexte d'analyse des transactions de ventes, et plus particulièrement, les transactions de vente stockées dans les bases de données des supermarchés (Agrawal et al., 1993). Il s'agissait d'analyser le comportement des clients lors de l'achat de différents produits dans un magasin, en identifiant des association/corrélations intéressantes en vue d'améliorer les *stratégies de marketing* pour augmenter les ventes et la gestion de l'information.

Selon la classification classique des techniques de *data mining* (i.e. descriptif et prédictif), l'extraction de règles d'association concerne surtout un *data mining* descriptif. Les règles d'association sont bien connues pour faciliter la compréhension des nouvelles connaissances par l'utilisateur, puisque des relations sous la forme SI (hypothèse) ALORS (conclusion) sont souvent considérées comme proches du raisonnement humain (Koskinen, 2012).

Cependant, cette méthode étant considérée comme un *data mining* descriptif, elle ne vise pas à trouver des caractéristiques générales relatives à une base de données, mais plutôt à identifier des règles d'association significatives qui expriment une corrélation entre un ensemble de valeurs d'attributs de la base de données.

### 1.5.3.2 Définition du problème

Le problème d'extraction de règles d'association concerne la recherche de motifs d'attributs (i.e. ensemble de valeurs d'attributs) fréquents dans une base de données. Considérons  $B = \{T, I\}$  une base de données, où  $T = \{t_1, t_2, \dots, t_n\}$  est un ensemble  $T$  de  $n$  **transactions** (ou objets) décrites par  $I = \{i_1, i_2, \dots, i_m\}$ , un ensemble  $I$  de  $m$  **items** (ou valeurs d'attributs), selon une relation binaire  $R \subseteq T * I$  entre  $T$  et  $I$ . Chaque transaction  $t_i$  dans  $B$  est donc constituée d'un sous-ensemble d'**items** contenus dans  $I$ .

Une règle d'association est formellement définie comme une relation entre deux **itemsets**<sup>3</sup> dans une base de données  $B$ . Elle représente les régularités de  $B$  à travers des relations de la forme « SI  $X$ , ALORS  $Y$  », notée  $X \rightarrow Y$ , où  $X$  et  $Y$  sont des sous-ensembles stricts de  $I$  ( $X, Y \in I$ ),  $X$  et  $Y$  sont mutuellement exclusifs ( $X \cap Y = \emptyset$ ), et la relation liant les deux **itemsets**  $X$  et  $Y$  est orientée de  $X$  vers  $Y$ .  $X$  est appelé **hypothèse** (ou prémisse) et  $Y$  est appelé **conclusion**.

Cette règle suggère qu'une forte relation existe entre  $X$  et  $Y$ , et peut être lue de la manière suivante : SI les **items** (valeurs d'attributs) contenus dans  $X$  existent dans une transaction, ALORS il est probable que les **items** contenus dans  $Y$  existent aussi dans la même transaction. Ainsi, la présence de  $X$  permet de conclure dans une certaine mesure sur la présence de  $Y$  pour l'ensemble des transactions analysées, d'où l'intérêt de ce type de connaissance.

---

<sup>3</sup> *Itemsets* : ensemble d'**items** (i.e. ensemble de valeurs d'attributs ou variables)

Une règle d'association est couramment définie par deux mesures ou paramètres classiques – le *support* et la *confiance*. Toutefois, nous allons présenter ci-dessous d'autres mesures qui sont aussi largement utilisées pour évaluer les règles dans ce contexte.

### 1.5.3.3 Des mesures pour évaluer les règles d'association

Les algorithmes d'extraction de règles d'association peuvent potentiellement générer un grand nombre de règles (dont beaucoup pourraient ne pas être intéressantes) suivant la taille et la dimension des bases de données analysées. Il existe de nombreuses mesures dans la littérature ayant pour objectif d'évaluer la qualité des règles d'un point de vue statistique. Nous présentons ci-dessous quelques exemples (voir l'Annexe 1 pour une liste plus exhaustive), les plus utilisés étant le support et la confiance qui sont à la base des algorithmes de recherche de règles.

#### Support

Le support (Agrawal et al., 1993) d'une règle d'association  $X \rightarrow Y$  est défini comme la proportion de transactions dans la base de données  $B$  qui contient à la fois  $X$  et  $Y$ . Il s'agit de la fréquence d'occurrence de la règle, qui varie dans l'intervalle  $[0,1]$  et ne prend pas en compte l'ordre des ensembles de valeurs d'attributs identifiés, i.e.  $Support(X \rightarrow Y) = Support(Y \rightarrow X)$ . Dans ce contexte, un *itemset* dont le support est supérieur ou égal à un seuil prédéfini *minsup* par l'utilisateur est appelé *itemset fréquent*.

$$Support(X \rightarrow Y) = P(X \cup Y) = P(XY) \quad (1)$$

#### Confiance

La mesure de confiance (Agrawal et al., 1993) indique la proportion de transactions dans  $B$  contenant  $Y$  parmi celles qui contiennent  $X$ , i.e. la probabilité conditionnelle de  $Y$  étant donné  $X$ . Cette mesure n'est pas symétrique (i.e. elle prend en compte l'ordre des *itemsets* identifiés) et varie aussi dans l'intervalle  $[0,1]$  ; de plus, elle ne prend pas en compte la taille de la base de données  $B$  ni la taille de la conclusion ( $P(Y)$ ).

$$Confiance(X \rightarrow Y) = P(Y|X) = \frac{P(X \cup Y)}{P(X)} = \frac{P(XY)}{P(X)} = \frac{Support(X \rightarrow Y)}{Support(X)} \quad (2)$$

#### Lift

La mesure *lift* (International Business Machines, 1996), appelée également « intérêt », prend en compte l'indépendance entre les *items* contenus dans  $X$  et dans  $Y$ , i.e.  $P(X) \times P(Y)$ . Il s'agit d'une mesure symétrique qui indique l'opportunité des transactions ayant  $X$  d'avoir  $Y$  qui varie dans l'intervalle  $[0, +\infty[$ . La mesure *lift* relie la confiance d'une règle au support de la conclusion ( $P(Y)$ ) de la règle (Giudici, 2003).

$$Lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X) \times P(Y)} = \frac{P(XY)}{P(X) \times P(Y)} \quad (3)$$

### Conviction

La conviction (Brin, Motwani, Ullman & Tsur, 1997) permet de mesurer pour chaque règle d'association la déviation de la dépendance entre la probabilité d'occurrence de l'hypothèse et la probabilité de non occurrence de la conclusion dans les transactions. Cette mesure n'est pas symétrique et varie dans l'intervalle  $[0, +\infty[$ .

$$Conviction(X \rightarrow Y) = \frac{P(X) \times P(\bar{Y})}{P(X \cup \bar{Y})} = \frac{P(X) \times P(\bar{Y})}{P(X\bar{Y})} \quad (4)$$

Quelle que soit la mesure statistique utilisée, les résultats issus du processus d'extraction doivent être interprétés soigneusement, en tenant compte du fait que les règles d'association extraites n'impliquent pas nécessairement un lien de causalité mais seulement la présence simultanée de deux *itemsets* dans des enregistrements.

#### I.5.3.4 Principaux algorithmes

Nous présentons ici les principaux algorithmes d'extraction de règles d'association, avec tout d'abord l'algorithme *Apriori* (Agrawal & Srikant, 1994), considéré comme l'algorithme traditionnel dans ce contexte, qui est le point de départ de nombreuses nouvelles propositions.

#### Algorithme *Apriori*

Le principe de l'algorithme *Apriori* (Agrawal & Srikant, 1994) réside dans le fait que si un ensemble d'*items* est « fréquent », alors tous ses sous-ensembles doivent aussi être fréquents. Ce principe permet de contrôler la croissance exponentielle des *itemsets* candidats.

Malgré les avancées des techniques d'extraction, le processus général adopté par la plupart des algorithmes d'extraction de règles d'association reste composé de deux sous-étapes : *i)* la *recherche d'itemsets fréquents* et, *ii)* la *génération des règles d'association*.

- **Recherche d'itemsets fréquents**

Le processus de génération d'*itemsets* fréquents dans l'algorithme *Apriori* est décrit dans l'Algorithme I.1. Soit *k-itemset* un ensemble d'*items* ayant *k items*,  $C_k$  un ensemble de *k-itemsets* candidats (*itemsets* potentiellement fréquents), et  $L_k$  un ensemble de *k-itemsets* fréquents (ceux avec le support minimum).

**Entrées :**

- $B$  : base de données transactionnelle
- $minsup$  : seuil minimal de support

**Sortie :**

- $\bigcup_k L_k$  : ensemble de tous les itemsets fréquents de  $B$

```

1)  $L_1 = 1\text{-itemsets}$  fréquents ;
2) pour ( $k = 2$  ;  $L_{k-1} \neq \emptyset$  ;  $k++$ ) faire
3)    $C_k = \text{Apriori-gen}(L_{k-1})$  ; // Génère l'ensemble de candidats  $C_k$  à partir de  $L_{k-1}$ 
4)   pour chaque transaction  $t \in B$  faire
5)      $C_t = \text{SousEnsemble}(C_k, t)$  ; // Sélection des candidats de  $C_k$  présents dans  $t$ 
6)     pour chaque candidat  $c \in C_t$  faire
7)        $count[c] = count[c] + 1$  ;
8)   fin pour
9)    $L_k = \{c \in C_k \mid count[c] \geq minsup\}$ 
10)   $k = k + 1$ 
11) fin pour
12) retourner  $\bigcup_k L_k$ 

```

*Apriori-gen* ( $L_{k-1}$ )

```

13) pour chaque itemset candidat  $c \in C_k$  faire
14)   pour chaque  $(k - 1)$ -sous-ensemble  $s$  de  $c$  faire
15)     si ( $s \notin L_{k-1}$ ) alors
16)       supprimer  $c$  de  $C_k$ 
17)   fin pour
18) fin pour

```

**Algorithme I.1. Découverte des *itemsets* fréquents dans l'algorithme *Apriori***

Comme décrit dans (Agrawal & Srikant, 1994), la première itération de l'algorithme (**ligne 1**) compte l'occurrence des *items* afin de déterminer les 1-*itemsets* (*itemsets* ayant 1 *item*) fréquents. Pendant les itérations suivantes, les *itemsets* fréquents  $L_{k-1}$  trouvés dans la  $(k - 1)^{\text{ème}}$  itération sont utilisés pour générer les nouveaux *itemsets* candidats  $C_k$ , à l'aide de la fonction *Apriori-gen* (**ligne 3**), qui comprend deux phases principales prenant comme argument  $L_{k-1}$  : union et élagage. Dans la phase d'union, tous les  $k$ -*itemsets* candidats potentiels sont générés à partir des  $(k - 1)$ -*itemsets* trouvés ( $L_{k-1}$ ). Puis, dans la phase d'élagage, tous les candidats générés dans la phase d'union qui ont certains  $(k - 1)$ -*itemsets* non-fréquents sont supprimés (Lazcorreta et al., 2008) : chaque *itemset*  $c \in C_k$  tel que certains  $(k - 1)$ -sous-ensembles de  $c$  ne sont pas dans  $L_{k-1}$  est supprimé de  $C_k$  (**lignes 13 à 18**). Ensuite, la base de données est analysée et les supports des candidats de  $C_k$  sont calculés (**lignes 4 à la 8**). La fonction *SousEnsemble* (**ligne 5**) reçoit l'ensemble de candidats et une transaction  $t$  de la base de données, et renvoie l'ensemble de candidats satisfaisant la transaction (**ligne 6 et 7**).

Le support de chaque candidat est alors pris en compte et est testé par rapport au seuil *minsup* (**ligne 9 et 10**). Finalement, l'algorithme se termine lorsque plus aucun *itemset* fréquent n'est généré, i.e.  $L_{k-1} = \emptyset$ .

- **Génération des règles d'association**

Le deuxième sous-problème est plus simple et concerne la génération des règles d'association satisfaisant le seuil de confiance requis par l'utilisateur (*minconf*) à partir des *itemsets* fréquents trouvés précédemment. Un algorithme intégré à l'algorithme *Apriori* est présenté dans l'Algorithme I.2.

**Entrées :**

- $L$  : ensemble des *itemsets* fréquents ( $\bigcup_k L_k$ )
- *minconf* : seuil minimal de confiance

**Sortie :**

- *Règles* : ensemble de règles d'association

- 1) **pour** chaque *itemset* fréquent  $l_k \in L, k \geq 2$  faire
- 2)       appel *gen-règles* ( $l_k, l_k$ ) ;
- 3) Procédure *gen-règles* ( $l_k : k$ -*itemset*,  $a_m : m$ -*itemset*)
- 4)  $A = \{(m-1)\text{-itemset } a_{m-1} \mid a_{m-1} \subset a_m\}$  ;
- 5) **pour** chaque  $a_{m-1} \in A$  faire
- 6)       *confidence*( $R$ ) = *support*( $l_k$ )/*support*( $a_{m-1}$ )
- 7)       **si** (*confidence*( $R$ )  $\geq$  *minconf*) **alors**
- 8)              $R = a_{m-1} \Rightarrow (l_k - a_{m-1})$
- 9)       **si** ( $m-1 > 1$ ) **alors**
- 10)           appel *gen-règles* ( $l_k, a_{m-1}$ )
- 11)           *Règles* = *Règles*  $\cup R$
- 12) **fin pour**
- 13) **retourner** *Règles*

**Algorithme I.2. Génération des règles dans l'algorithme *Apriori* (Agrawal & Srikant, 1994)**

Soit  $L$  l'ensemble d'*itemsets* fréquents et  $l_k \in L$ . L'algorithme considère successivement chaque *itemset* fréquent de  $L$  de taille supérieure à 1 (**lignes 1 à 13**). Pour chacun de ces *itemsets*  $l_k$ , l'algorithme trouve tous les sous-ensembles  $a_{m-1}$  de  $l_k$  (**ligne 4**) et propose un ensemble de règles candidates sous la forme  $a_{m-1} \rightarrow (l_k - a_{m-1})$ , qui doivent être vérifiées en tenant compte du *minconf* (**lignes 6 à 8**). Les règles générées à partir de  $l_k$  sont les règles dont l'union de l'hypothèse et de la conclusion donne l'*itemset*  $l_k$ . La procédure *gen-règles* est alors appelée (**ligne 10**) pour insérer dans « *Règles* » les règles valides générées à partir de  $l_k$  dont la conclusion contient plus d'un *item*.

Des améliorations ou extensions de l'algorithme *Apriori* ont notamment été suggérées dans la littérature afin de diminuer le temps d'exécution, la consommation de mémoire ou encore d'améliorer son efficacité (Chen et al., 2008) (Yu & Zhou, 2010).

Une analyse des différentes propositions développées dans la littérature pour améliorer ce processus d'extraction de règles d'association, surtout lors de la première étape du processus (considérée comme la partie plus complexe du fait du problème exponentiel de la génération des *itemsets* fréquents), est par exemple présentée dans (Marinica, 2010).

Nous décrivons ci-dessous les principales approches abordant cette problématique.

### Techniques alternatives pour la génération des candidats

Dans la plupart des algorithmes de génération de candidats, y compris l'algorithme *Apriori*, des contraintes sont incorporées afin de valider un ensemble d'*itemsets* candidats identifiés. Les premiers algorithmes incorporant des contraintes ont été AIS (Agrawal et al., 1993) et SETM (Houtsma & Swami, 1993) ; cependant les améliorations apportées n'ont pas été significatives par rapport au problème de l'explosion d'*itemsets* fréquents.

C'est pourquoi d'autres algorithmes basés sur *Apriori* ont été développés avec le même objectif, i.e. optimiser la génération d'*itemsets* fréquents. Des améliorations ont été proposées dans la littérature en introduisant d'autres contraintes dans le processus, par exemple des représentations condensées, des partitionnements de l'ensemble de données, l'élagage d'un ensemble de données ou encore la réduction de l'accès aux données (Marinica, 2010). Ci-dessous, nous présentons quelques exemples.

- L'algorithme **Apriori-TID** (Agrawal & Srikant, 1994) est une extension d'*Apriori* qui diminue le nombre d'itérations à parcourir au cours de l'analyse de la base de données, en utilisant un identifiant pour chaque *itemset* (TID).
- L'algorithme **Direct Hashing and Pruning** (DHP) (Park et al., 1997) utilise des tables de hachage afin de réduire le nombre de candidats générés.
- L'algorithme **Partition** (Savasere et al., 1995) propose un partitionnement de la base de données afin d'optimiser la consommation de mémoire.
- **Dynamic Itemset Counting** (DIC) (Brin, Motwani, Ullman & Tsur, 1997) a introduit l'idée de candidats  $(k + 1)$ , i.e. quand un  $k$ -*itemset* est considéré comme fréquent, tous les  $(k + 1)$ -*itemsets* candidats (que  $k$ -*itemset* peut produire) sont générés.
- L'algorithme **Tree Projection** (Agarwal et al., 2001) utilise l'arbre d'un ensemble d'énumérations pour générer des candidats.

D'autres algorithmes ont introduit par exemple des hyper structures de données compactes **FP-tree** (*Frequent Pattern tree*). Il s'agit des algorithmes *Pattern-Growth* qui contiennent une liste d'items et des modèles de frames pour le stockage des données dans la génération des *itemsets* fréquents. L'algorithme **FP-Growth** (Han & Pei, 2000) permet en effet de découvrir les *itemsets* fréquents ( $L$ ) sans la génération des candidats ( $C$ ), étant donné que la génération de candidats consomme

mémoire et temps dans l'algorithme *Apriori*. La première étape de cet algorithme concerne la création d'une structure de stockage de données, appelé *FP-tree*. L'extraction des *itemsets* fréquents est ensuite effectuée directement à partir de cette structure. Il s'agit ainsi d'une alternative qui semble intéressante pour générer les *itemsets* fréquents sans générer d'abord des candidats.

Une alternative concerne les algorithmes basés sur des représentations condensées. Dans cette classification, nous considérons les algorithmes basés sur l'analyse de concepts formels (*Formal Concept Analysis - FCA*) (Zaki & Ogihara, 1998) afin d'extraire des *itemsets* fermés fréquents (*frequent closed itemsets*). Différents formalismes ont été proposés dans la littérature dans ce contexte : l'algorithme **Close** (Pasquier et al., 1999), l'algorithme **PASCAL** (Bastide et al., 2000), l'algorithme **CLOSET** (FP-tree) (Pei et al., 2000), et l'algorithme **CHARM** (Zaki & Hsiao, 2002) en sont des exemples.

### Techniques alternatives pour la génération de règles d'association

Dans cette section nous allons aborder des améliorations concernant la génération des règles d'association, i.e. des solutions proposées pour réduire le nombre de règles extraites par les techniques classiques. Il s'agit d'algorithmes basés sur diverses contraintes (liées par exemple aux connaissances de l'utilisateur, aux connaissances du domaine, aux contraintes sur les données, aux contraintes dimensionnelles, aux contraintes sur les règles, etc.) afin d'extraire seulement les règles qui satisfont ces contraintes. Dans le cas de l'algorithme *Apriori* et de ses variantes, les deux contraintes classiques utilisées sont le *minsup* et le *minconf*.

D'autres algorithmes se sont plutôt intéressés aux techniques de réduction des règles redondantes. Dans ce contexte, les premières idées concernant la génération de règles de non-redondance ont été suggérées dans (Zaki, 2000). Dans (Pasquier et al., 2005) les auteurs ont proposé deux approches pour représenter les règles d'association redondantes afin de n'extraire que les règles non-redondantes. Finalement, d'autres techniques ont été aussi implémentées dans (Ashrafi et al., 2007) pour aborder cette problématique.

#### 1.5.3.5 Domaines d'application des règles d'association

Plusieurs applications réelles des systèmes d'ECD utilisant l'extraction de règles d'association ont été mises en œuvre dans différents domaines, notamment dans la planification commerciale, les télécommunications, le domaine médical, la maintenance industrielle, la sécurité, la productivité ou le domaine de chimique, entre autres (Pasquier, 2000).

Nous présentons ci-dessous des exemples dans différents domaines.

- L'analyse des paniers de consommation (Agrawal et al., 1993) (Fayyad et al., 1996), abordé précédemment, est l'exemple classique de la technique d'extraction de règles d'association.
- Les données stockées dans les bases de données médicales (i.e. hôpitaux, laboratoires d'analyse, cabinets médicaux, ...) sont en train d'augmenter d'une manière significative, d'où l'intérêt particulier d'analyser les relations entre les différents attributs afin d'améliorer les processus de prise de décision et la *gestion médicale*. Une étude sur l'extraction de règles

d'association dans des bases de données médicales a été présentée dans (Delgado et al., 2001) visant l'amélioration des tâches de gestion des patients.

- Les règles d'association dans le *domaine de la maintenance* peuvent faciliter la compréhension du comportement des équipements afin de détecter par exemple les défaillances potentielles des systèmes de production à partir des données collectées (Young et al., 2010). Diverses applications montrent en effet que la découverte de règles d'association dans les bases de données de maintenance est essentielle afin de valoriser et d'utiliser efficacement ces données stockées au quotidien. Nous nous intéressons au potentiel de ce domaine d'application qui sera abordé plus en détail lors du Chapitre IV.
- Pour améliorer la *qualité des produits*, des règles d'association ont été utilisées par exemple pour détecter des machines fabriquant des produits défectueux dans (Chen et al., 2005). Dans (Kusiak, 2002), un algorithme de *data mining* extrait des associations entre les paramètres de contrôle et la qualité des produits dans le processus de formage des métaux.
- A partir des bases de données de production, les auteurs ont exposé dans (Shahbaz et al., 2006) l'application de règles d'association pour extraire des informations utiles suggérant des améliorations dans la *conception de produits* et dans les *procédés de fabrication*. D'autres exemples traitant des règles d'association dans la conception de produits ont été suggérés dans (Jiao et al., 2008) (Liao et al., 2008) (Tsai et al., 2003).
- Concernant la *configuration des systèmes* de production, Chen (Chen, 2003) a utilisé une approche basée sur l'extraction de règles pour résoudre des problèmes de configuration des cellules dans le système. En effet, des associations entre les machines sont découvertes à partir de la base de données du processus.
- A propos de la gestion de la relation clients (*Customer Relationship Management* - CRM) dans les entreprises, des règles d'association peuvent être générées pour prédire les mouvements ou comportement potentiels des clients d'une entreprise à l'autre, permettant aux entreprises d'être plus proactives (Morita et al., 2000).

## I.6 BILAN

Dans ce premier chapitre, nous avons abordé les principales sources de connaissances au sein des entreprises, i.e. l'extraction de connaissances auprès des experts, la capitalisation et la formalisation des expériences passées et l'analyse des bases de données disponibles pour extraire de nouvelles connaissances. Toutefois, notons qu'il est possible de combiner plusieurs de ces techniques pour construire une mémoire d'entreprise plus robuste.

Plus particulièrement, nous avons étudié de manière séparée deux courants scientifiques pouvant être des sources de connaissances possibles par rapport à notre problématique : la démarche de Retour d'Expérience (REx) et les techniques d'Extraction de Connaissances à partir de Données (ECD).



Les différents types de retour d'expérience qui ont été abordés nous permettent de mieux situer le type de retour d'expérience auquel nous nous intéressons. Nos travaux considèrent en effet un REx de type hybride qui combine plusieurs types de REx et qui présente les caractéristiques suivantes :

- REx positif/négatif : nous nous intéressons à un REx positif (permettant la capitalisation des informations issues des événements positifs ou bonnes pratiques), ainsi qu'à un REx négatif (normalement plus facile à mettre en œuvre en considérant les situations de résolution de problèmes).
- REx cognitif/statistique : nous nous intéressons à une démarche hybride de retour d'expérience en tenant compte *i)* de la mise sous une forme spécifique des expériences recueillies à l'aide d'un formalisme de représentation de connaissances afin de faciliter leur future réutilisation, et *ii)* de l'application de méthodes relatives à la fouille de données (ou techniques de *data mining*) afin d'extraire de nouvelles connaissances qui doivent à notre avis nécessairement être évaluées et validées par les experts du domaine avant leur réutilisation et/ou leur intégration dans le système industriel. En effet, les informations de terrain initialement capitalisées (expériences recueillies) restent fragmentaires, et nécessitent un effort pour être généralisées (d'où l'intérêt des techniques d'extraction de connaissances). Dans notre contexte, le principal objectif de l'extraction de règles d'association est de transformer ces informations disponibles issues des expériences passées stockées dans des bases de données, en connaissances plus génériques afin d'aider la prise de décision dans les entreprises.

Ainsi, sur la base de ces caractéristiques, nous allons présenter dans le **Chapitre II** une démarche conceptuelle complète de retour d'expérience intégrant aspects cognitifs et statistiques, basée sur une architecture de mémoire d'entreprise composée d'une base d'expériences et d'une base de connaissances sous la forme de règles. Pour la structuration de cette mémoire d'entreprise, un formalisme de la représentation des connaissances est un élément clé, facilitant le stockage, le traitement, la compréhension par l'utilisateur et une réutilisation future des différents types de connaissances modélisées. Nous allons présenter notre choix sur ce formalisme de représentation dans la suite.



## II COUPLAGE REx-ECD : UNE DEMARCHE ORIGINALE DE GESTION DES CONNAISSANCES

---

II.1	POSITIONNEMENT DE LA DEMARCHE DE RECHERCHE .....	38
II.1.1	POSITIONNEMENT .....	38
II.1.2	INTERETS D'UN COUPLAGE REx-ECD .....	40
II.2	FORMALISATION DES CONNAISSANCES A L'AIDE DES GRAPHS CONCEPTUELS.....	41
II.2.1	GENERALITES .....	42
II.2.2	VOCABULAIRE DU DOMAINE.....	42
II.2.3	NOTION DE GRAPHE CONCEPTUEL .....	44
II.2.4	SUBSOMPTION ET HOMOMORPHISME : DES OUTILS DE RAISONNEMENT.....	44
II.2.5	GRAPHS CONCEPTUELS EMBOITES .....	46
II.2.6	INTERET DES GCs DANS NOTRE APPROCHE REx-ECD.....	46
II.3	DESCRIPTION DE LA DEMARCHE GLOBALE REx-ECD.....	47
II.3.1	PRETRAITEMENT DE L'ECD ASSOCIE A LA PHASE DE CAPITALISATION DU REx .....	48
II.3.2	TECHNIQUES DE DATA MINING ASSOCIEES A LA PHASE DE TRAITEMENT DU REx.....	50
II.3.3	POST-TRAITEMENT DE L'ECD ASSOCIE A LA PHASE D'EXPLOITATION DU REx.....	59
II.4	BILAN .....	62

Au cours du premier chapitre, nous avons présenté un état de l'art concernant les généralités de la mise en œuvre d'une démarche de Retour d'Expérience (REx) au sein des organisations, ainsi que des techniques complémentaires qui sont associées à ce processus pour exploiter les informations disponibles (techniques d'Extraction de Connaissances à partir de Données (ECD)).

Nous avons exposé l'intérêt de ces deux démarches (REx et ECD) de manière séparée. Dans ce chapitre, l'objectif est de présenter la mise en œuvre d'une démarche originale de gestion des connaissances issues des expériences, en tenant compte de la complémentarité entre ces deux approches émergentes.

Pour cela, nous présentons dans un premier temps le positionnement de notre démarche de recherche et l'intérêt d'un couplage REx-ECD dans les entreprises. Ensuite, nous exposons le formalisme de représentation de connaissances proposé pour supporter la méthodologie générale, i.e. les Graphes Conceptuels (GCs). Finalement, nous présentons les grandes lignes de l'approche REx-ECD dans un contexte industriel et les principales étapes mises en œuvre afin de :

- Structurer une base d'expériences formalisées, concernant notamment la représentation du vocabulaire du domaine d'application et des informations qui caractérisent les expériences recueillies (connaissances spécifiques) issues des processus industriels. Le principal objectif de cette base d'expériences est de faciliter le traitement, le partage et la réutilisation future de ces expériences, par exemple dans la résolution d'un nouveau problème à l'aide des mécanismes de raisonnement ;
- Générer une base de règles extraites (connaissances plus génériques) à partir d'un processus d'extraction de nouvelles connaissances, et plus particulièrement, d'extraction de règles d'association. Ces connaissances (sous forme de règles), plus génériques que les connaissances contenues dans les expériences, sont tirées d'informations répétitives dans les bases de données disponibles, et doivent aider la prise de décisions stratégiques et l'amélioration continue des processus.

## II.1 POSITIONNEMENT DE LA DEMARCHE DE RECHERCHE

### II.1.1 POSITIONNEMENT

Au sein des organisations, de quantités de données considérables, issues des événements passés, sont recueillies et stockées dans des bases de données, des data warehouses ou d'autres dépôts d'information (Chen et al., 1996) (Han & Kamber, 2006) (Liu & Motoda, 1998). Les problèmes potentiels liés à l'explosion des données dans différents domaines et secteurs ont attiré une attention considérable de la part des chercheurs en sciences de l'information comme des décideurs au sein des entreprises (Chen & Zhang, 2014), et des solutions ont été recherchées pour analyser ces importants volumes de données (appelées souvent « *Big Data* ») et en tirer de nouvelles connaissances.

Les démarches classiques de Retour d'Expérience (REx) sont nées du besoin de capitaliser les connaissances attachées aux expériences passées dans les entreprises. Plus particulièrement, le retour d'expérience de type cognitif concerne une approche émergente qui rend disponible des informations significatives issues des expériences (Béler, 2008) (Jabrouni, 2012) (Kamsu-Foguem et al., 2008) (Rakoto, 2004) (Weber et al., 2001) (Weber & Aha, 2003). D'autre part, les techniques

d'Extraction de Connaissances à partir de Données (ECD) sont nées du besoin de valoriser les bases de données disponibles dont la taille croît de manière exponentielle (Alzghoul & Löfstrand, 2011) (Fayyad et al., 1996) (Han & Kamber, 2006) (Kargupta et al., 2008) (Nisbet et al., 2009) (Poncelet et al., 2008).

Nous proposons une démarche de gestion des connaissances en tenant compte de ces deux approches. Nous nous intéressons en effet à la capitalisation et à la formalisation des informations issues des expériences, mais aussi à la création de connaissances à partir de l'analyse de cette information disponible.

### **Pourquoi le REx ?**

Comme nous l'avons déjà indiqué dans le chapitre précédent, capturer et stocker des expériences passées ou des connaissances spécifiques est considéré comme étant plus facile que capturer et stocker directement des connaissances génériques hors contexte. En effet, les expériences qui se situent dans un contexte de résolution de problème, ou dans une situation particulière, sont plus facilement observables et descriptibles par les groupes humains (Bergmann, 2002) lors de leurs activités quotidiennes, et ont un fort potentiel comme source de nouvelles connaissances.

### **Pourquoi extraire des nouvelles connaissances ?**

Étant donné que les quantités de données stockées dans les entreprises augmentent d'une manière exponentielle, leur exploitation devient d'un grand intérêt. Concernant le défi de traiter de grands volumes de données, il n'y a aucun doute que les entreprises vont davantage exploiter les « *Big Data* » dans l'avenir (Chen & Zhang, 2014), pour créer des avantages concurrentiels.

Le principal objectif dans l'analyse des bases de données existantes est de valoriser le capital intellectuel des entreprises, d'aider la prise de décision et d'améliorer les performances des processus en fournissant aux acteurs du domaine une information concise et structurée (nouvelles connaissances) décrivant le contenu des bases de données analysées. A cette fin, nous nous intéressons ici à la découverte de connaissances générales à partir de l'information caractérisant les expériences passées ; il s'agit d'un processus qui se trouve hors de la portée principale des approches classiques de gestion des expériences (Bergmann, 2002).

### **Pourquoi l'extraction de règles d'association ?**

Parmi les techniques de *data mining* étudiées, nous nous intéressons dans notre contexte d'étude à l'extraction de règles d'association principalement pour les deux raisons présentées dans (Marinica, 2010) : *i)* le modèle des motifs extraits est relativement simple et compréhensible par un utilisateur non spécialiste en *data mining*, et *ii)* la charge de travail des utilisateurs pendant le processus d'extraction (notamment dans la phase de *data mining*) reste limitée. En somme, le principal objectif des règles d'association dans notre travail est de maîtriser dynamiquement le lien entre les différents attributs analysés décrivant les expériences, valorisant ainsi l'information disponible.

### II.1.2 INTERETS D'UN COUPLAGE REx-ECD

Parmi les différentes méthodes du domaine de la gestion des connaissances permettant la définition d'une mémoire d'entreprise, nous nous sommes focalisés notamment sur les méthodes de gestion des expériences et sur les méthodes d'extraction de connaissances à partir de données (Ben Sta, 2006). C'est pourquoi nous proposons dans ce chapitre de prendre en compte ces deux méthodes en parallèle afin de concevoir une mémoire d'entreprise (appelée base de REx-ECD par la suite) et d'allier leurs avantages pour créer, capitaliser et exploiter des connaissances issues de l'expérience dans un contexte industriel.

Le couplage REx-ECD donne ainsi lieu à une approche plus générique d'aide à la décision à partir des informations caractérisant les expériences passées. En effet, les approches d'ECD étant considérées comme complémentaires à la démarche classique du REx (Rakoto, 2004), elles permettent de générer de nouvelles connaissances par l'analyse des événements passés, i.e. par la généralisation du résultat de nombreuses expériences pour constituer et améliorer un référentiel de connaissances globales des entreprises.

L'approche couplée proposée présente plusieurs aspects originaux visant à répondre à certaines limitations des approches classiques trouvées dans la littérature :

- Le processus de retour d'expérience intègre des activités et techniques d'extraction de connaissances à partir de données (processus de *data mining*) afin de générer de nouvelles connaissances plus génériques liées aux expériences.
- Les techniques de *data mining*, au niveau scientifique, se situent à l'intersection des méthodes statistiques, de l'informatique et de l'intelligence computationnelle. Le but de l'activité de *data mining* choisie dans notre approche n'est pas de résoudre un problème spécifique, mais au contraire d'explorer tous les liens possibles entre les ensembles de valeurs d'attributs identifiés par un algorithme d'extraction de règles, et d'analyser leur utilité possible pour l'amélioration des processus industriels.
- Puisque différents types de règles d'association peuvent ainsi être obtenues, nous pensons qu'il est intéressant de présenter à l'utilisateur final (expert du domaine) les relations existantes entre les types de règles identifiées, modélisées à l'aide d'un formalisme de représentation de connaissances (un de nos objectifs étant de fournir des outils permettant d'analyser la structure et la sémantique des connaissances produites).
- Pour cette même raison, il devient intéressant de fournir aussi à l'utilisateur final un outil comprenant des mécanismes permettant de filtrer l'ensemble de règles extraites selon ses différents points de vue et/ou ses attentes dans une situation particulière.
- Finalement, des conclusions intéressantes peuvent être tirées non seulement des règles qui sont présentes après le processus d'extraction, mais aussi du fait que certaines règles attendues sont absentes. Nous allons montrer que le fait de comparer les règles présentes et celles absentes peut aider à identifier des anomalies dans la façon avec laquelle une expérience est structurée dans la base de données, ou peut servir à définir plus précisément les

caractéristiques opérationnelles des processus dans l'entreprise. À notre connaissance, il n'existe pas de travaux dans la littérature portant sur ce sujet.

## II.2 FORMALISATION DES CONNAISSANCES A L'AIDE DES GRAPHES CONCEPTUELS

La mise en œuvre d'un processus de retour d'expérience lié à un processus d'extraction de connaissances sous la forme de règles requiert un formalisme de représentation de connaissances afin de structurer et de modéliser la base de REx-ECD, puis de faciliter leur traitement et leur exploitation future.

Parmi les formalismes de représentation de connaissances exposés dans le Chapitre I (section I.4), nous adoptons celui des Graphes Conceptuels (GCs) qui nous paraît le plus adapté à notre contexte d'étude, car il fournit des caractéristiques intéressantes tant en termes d'expressivité que d'inférence (avec ses mécanismes de raisonnement). Les GCs seront ainsi utilisés afin de structurer d'une manière formelle et compréhensible les différents types de connaissances dans une représentation unifiée de la connaissance.

En ce qui concerne l'expression de la connaissance et les mécanismes de raisonnement graphiques (Chein et al., 2013), les approches basées sur les graphes présentent plusieurs avantages. Les principaux avantages des GCs sont présentés ci-dessous :

- Il s'agit d'un formalisme de représentation - et aussi de raisonnement - qui reste actuellement l'unique modèle basé sur la logique ayant une interprétation équivalente dans la théorie de graphes (Thomopoulos et al., 2010).
- La représentation des connaissances est entièrement graphique et est proche d'une expression en langage naturel. De même, le raisonnement repose sur des opérations de graphes, facilitant ainsi sa compréhension intuitive et son interprétation par l'utilisateur non-expert. Ces opérations de graphes agissent directement sur les connaissances représentées, sans nécessiter un langage logique, ce qui évite notamment le problème de la subsomption présenté dans les logiques de description (LDs) (McGuinness & Borgida, 1995).
- Les « *mappages* », ou transformations, du langage naturel aux GCs ont été étudiés dans la littérature et sont assez prometteurs, par exemple pour la modélisation des connaissances procédurales (Lai, 2007).
- Les GCs permettent de représenter différents types de connaissances (descriptions factuelles, modèles de connaissances, règles d'inférence et/ou contraintes par exemple (Baget & Mugnier, 2002)), et permettent également de structurer et contextualiser les connaissances à travers les graphes emboîtés (*nesting of graphs* en anglais) (Chein & Mugnier, 2008).
- La représentation formelle des connaissances à l'aide des GCs capture la sémantique de la connaissance et permet de raisonner sur cette connaissance (Lai, 2007).

- Les GCs fournissent des outils de raisonnement qui facilitent la visualisation et la vérification de la connaissance modélisée par les utilisateurs finaux (Dieng-Kuntz & Corby, 2005).
- En outre, les GCs peuvent être traduits dans d'autres formalismes de représentation de connaissances, tels que les langages de web sémantique (Sowa, 2000) (Yao & Etzkorn, 2006) ou les réseaux de Petri (Sowa & Zachman, 1992) par exemple.

### II.2.1 GENERALITES

Le formalisme des GCs peut être considéré comme un compromis entre les langages graphiques et les langages formels, puisqu'il présente des qualités visuelles mais qu'il est aussi fondé sur la logique (Chein & Mugnier, 2008). En effet, ce formalisme a une syntaxe bien définie et une sémantique formelle permettant de raisonner à partir de ses représentations (Kamsu-Foguem et al., 2013).

Nous présentons dans la suite les principales caractéristiques de ce formalisme de représentation des connaissances ainsi que ses principales techniques de raisonnement.

### II.2.2 VOCABULAIRE DU DOMAINE

Etant donné que la plupart des problèmes industriels implique une communication entre experts de différents domaines, une attention particulière a été portée sur les ontologies, assurant que l'information/connaissance échangée par les différents acteurs est significative et interprétée de la même manière (Uschold & Grüninger, 1996). Nous accordons ici un intérêt particulier aux ontologies.

Une ontologie est définie dans (Gruber, 1995) comme une spécification explicite et formelle d'une conceptualisation partagée. La notion de « conceptualisation » fait référence à un modèle abstrait du domaine concerné. Le fait que la conceptualisation doit être « partagée » implique une connaissance consensuelle, acceptée par le groupe. « Spécification explicite » suppose que les concepts identifiés et les contraintes qui les lient sont explicitement définis. L'aspect « formel » permet de garantir que l'ontologie est lisible par une machine, i.e. « computationnelle » (Studer et al., 1998).

Dans ce contexte, les graphes conceptuels permettent de décrire différents types de connaissances, e.g. assertions, requêtes, règles, contraintes, etc. Dans ce chapitre, ils sont utilisés principalement pour représenter des assertions (*les expériences passées et les connaissances extraites*), tandis que dans le chapitre suivant, ils sont utilisés pour exprimer les requêtes de l'utilisateur du domaine d'application sur ces connaissances formalisées.

Pour représenter ces assertions, des entités ou objets (appelés « concepts » dans la communauté des GCs) sont nécessaires, et sont liés par des relations. Un GC est ainsi composé de deux types de nœuds : les nœuds « concept » représentant les entités et les nœuds « relation » représentant les liaisons entre ces entités.

Pour la représentation des connaissances à l'aide des GCs, Chein et Mugnier (Chein & Mugnier, 1992) ont proposé tout d'abord la définition d'un vocabulaire de base (appelé support) qui fournit les



éléments essentiels pour le vocabulaire du domaine, i.e. une description partagée des concepts et des relations permettant une description de la connaissance du domaine.

Ce support va régir l'ensemble des GCs et peut être considéré comme une ontologie rudimentaire (Chein & Mugnier, 2008) : il s'agit d'une description de la connaissance au niveau conceptuel, précisant le vocabulaire du domaine et la sémantique qui lui est attachée (Fürst & Trichet, 2009).

Le support ( $S$ ) dans les GCs, ou connaissance ontologique fondamentale, est défini comme un triplet  $(TC, TR, I)$  ; l'ensemble fini et ordonné des types de concepts est noté  $TC$  et est appelé une « hiérarchie des types de concepts », tandis que l'ensemble des relations est noté  $TR$  et représente une « hiérarchie des types de relations » (Chein & Mugnier, 2008), traduisant ainsi la connaissance déclarative. En outre, un concept individuel ( $c$ ) est composé d'un « type » (étiquette) et d'un « marqueur » :  $[<type> : <marqueur>]$ . Ces marqueurs peuvent être individuels ou génériques : un marqueur individuel appartient à un ensemble de marqueurs individuels  $I$  (qui désigne une instance considérée, une constante ou une variable), et un marqueur générique « \* » indique une instance non définie (Fürst & Trichet, 2009). De ce fait, le  $type(c) \in TC$  et le  $marqueur(c) \in I \cup \{*\}$ .

Nous illustrons sur la Figure II.1 un exemple de support (en tenant compte de  $TC$  et de  $TR$ ) dans le domaine de la maintenance (Kamsu-Foguem & Noyes, 2013). Dans  $TC$  (partie gauche de la Figure II.1), les principaux concepts du domaine sont représentés (i.e. « objet » concernant les différents objets du système, « état » décrivant les états du service, « maintenance » décrivant les type d'intervention, « indicateur » concernant des attributs pour évaluer la sûreté de fonctionnement, « entrave » décrivant les circonstances qui peuvent affecter la sûreté de fonctionnement, « durée » concernant le temps de l'intervention ou d'arrêts). D'autre part, dans  $TR$  (partie droite de la Figure II.1) sont présentées les relations nécessaires pour relier les concepts (i.e. « relation temporelle », « relation usuelle », « relation logique », « relation spatiale »).

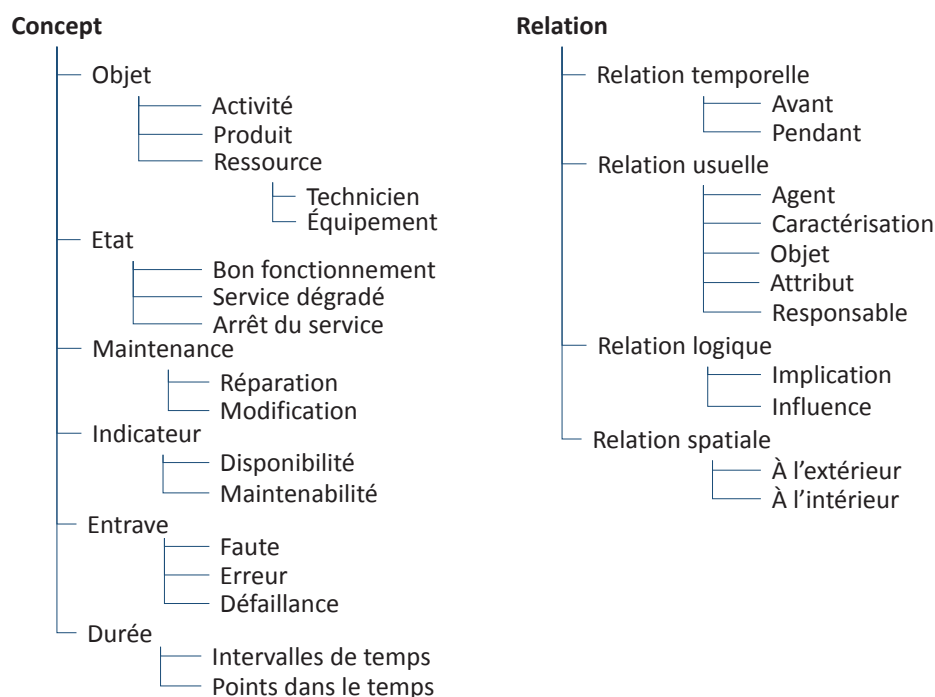


Figure II.1 : Exemple de support (adaptée de (Kamsu-Foguem & Noyes, 2013))

### II.2.3 NOTION DE GRAPHE CONCEPTUEL

Un graphe conceptuel est défini comme un graphe fini, orienté, connexe et biparti, qui comporte deux types de nœuds : des nœuds concepts (indiqués par des rectangles), et des nœuds relations (indiqués par des ovales), permettant ainsi de représenter dans notre contexte une connaissance procédurale. En notation linéaire, les nœuds concept sont écrits entre []-crochets tandis que les nœuds relations sont écrits entre ()-parenthèses.

Dans la Figure II.2, nous présentons un graphe conceptuel, noté  $GC-H$ , défini à partir du support présenté dans la Figure II.1. Ce graphe est composé de quatre nœuds concept (« défaillance » concernant un marqueur individuel « court-circuit\_09 », « arrêt du service », « durée » et « disponibilité ») et de trois relations binaires (« agent », « caractérisation » et « influence »). En ce qui concerne le type de concept « défaillance », on désigne, par exemple, une défaillance non précisément identifiée par le concept générique [Défaillance : \*], et une défaillance concernant un court-circuit par le concept individuel [Défaillance : court-circuit].

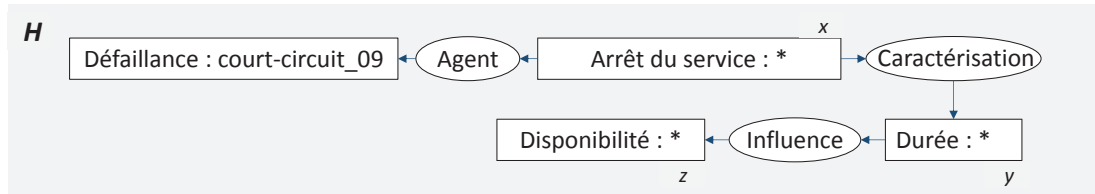


Figure II.2 : Exemple de GC (Kamsu-Foguem & Noyes, 2013)

Le graphe conceptuel  $H$  se lit de la manière suivante, en tenant compte du sens des arcs : *l'arrêt d'un service est agent d'une défaillance concernant un court-circuit (09). Cet arrêt est caractérisé par une durée, qui aura une influence sur la disponibilité du service.*

D'autre part, chaque GC possède également une représentation équivalente exprimée en logique des prédicats (Sowa, 1984). L'interprétation logique du graphe conceptuel  $H$  de la Figure II.2 est la suivante :

$$F(H): \exists x, y, z (Défaillance(court - circuit\_09) \wedge Arrêt\ du\ service(x) \wedge Durée(y) \wedge Disponibilité(z) \wedge Agent(x, court - circuit\_09) \wedge Caractérisation(x, y) \wedge Influence(y, z))$$

(Kamsu-Foguem et al., 2013).

### II.2.4 SUBSOMPTION ET HOMOMORPHISME : DES OUTILS DE RAISONNEMENT

L'enjeu des formalismes de représentation des connaissances ne porte pas uniquement sur leur capacité à représenter les connaissances, mais aussi sur les mécanismes permettant de raisonner sur ces descriptions (Stuber, 2007). Les raisonnements basés sur la théorie des graphes dans les GCs concernent la relation de *subsumption*, i.e. des opérations de graphes basées essentiellement sur les homomorphismes des graphes (Hell & Nesetril, 2004).

Une relation de *spécialisation/généralisation* ( $\leq$  ou  $\geq$ ) - ou relation de subsomption - est le principe essentiel pour le raisonnement avec les GCs simples (Mugnier, 1995) (Mugnier, 2000) en se basant sur les types de concepts et de relations définis dans le support ontologique. Étant donnés, par exemple, deux graphes  $G$  et  $H$  sur le même support (Figure II.3),  $H \leq G$  ( $G$  est dit « subsumer »  $H$ ) si  $H$  peut être obtenu à partir de  $G$  par une opération globale, qui est essentiellement un homomorphisme de graphe. Ainsi,  $H \leq G$  implique l'existence d'un homomorphisme de  $G$  vers  $H$  (Mugnier, 2000). Plus particulièrement,  $H$  est une spécialisation de  $G$  ( $H \leq G$ ) si  $H$  peut être dérivé de  $G$  par des opérations de spécialisation, ou  $G$  est une généralisation de  $H$  ( $G \geq H$ ), si  $G$  peut être dérivé de  $H$  par des opérations de généralisation (Chein & Mugnier, 2008).

Dans la communauté des GCs, un homomorphisme est traditionnellement appelé une **projection** (Chein & Mugnier, 1992), et permet un calcul effectif de la relation d'ordre  $\leq$ . Ainsi, l'existence d'une telle projection d'un GC- $G$  vers un GC- $H$  décrit un mappage des concepts de  $G$  vers les concepts de  $H$ , et des relations de  $G$  vers les relations de  $H$ . Cela signifie que la connaissance représentée par  $H$  est déductible de la connaissance représentée par  $G$  en tenant compte du support ( $TC$  et  $TR$ ). Afin d'illustrer cette opération de projection, compte tenu du support de la Figure II.1, nous présentons ci-dessous les graphes conceptuels  $G$  et  $H$  et la relation de subsomption existant entre eux (Figure II.3).

Le graphe  $G$  se lit de la manière suivante : *l'état d'un service est agent d'une entrave non précisément identifiée. Cet état a une durée non précisée, qui aura une influence sur un indicateur (attribut de la sûreté de fonctionnement) du système.*

En ce qui concerne la relation de subsomption entre  $G$  et  $H$ , par exemple, le type de concept « Défaillance » dans le graphe  $H$  concerne une spécialisation (ou un sous-concept) du concept « Entrave » dans le graphe  $G$  en tenant compte de  $TC$  (Figure II.1) ; autrement dit, le concept « Entrave » est une généralisation (ou sur-concept) du type de concept « Défaillance ». Par rapport aux relations, le principe est le même ; ainsi, par exemple, la relation « Caractérisation » dans le graphe  $H$  est une spécialisation de la relation « Relation usuelle » dans  $G$ , et vice-versa.

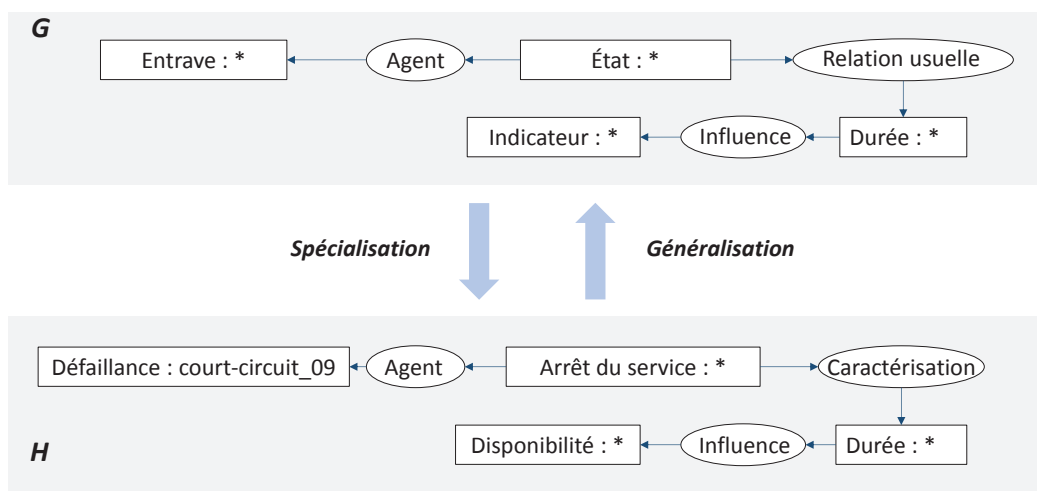


Figure II.3. Relation de subsomption (Kamsu-Foguem & Noyes, 2013)

## II.2.5 GRAPHES CONCEPTUELS EMBOITES

Les GCs emboîtés sont une extension des GCs simples, capable de représenter les notions d'information interne et externe (Chein & Mugnier, 2008). Un graphe conceptuel emboîté est tel que les nœuds concepts peuvent inclure d'autres graphes représentant une description partielle. Un nœud concept peut donc représenter une information externe (i.e. un contexte), et une description partielle peut être considérée comme une information interne au contexte. Il s'agit donc de l'incorporation de graphes dans d'autres graphes, pour permettre de mieux représenter les connaissances.

Pour illustrer ceci, considérons le premier graphe de la Figure II.4, exprimant que « *le technicien Jean est le responsable d'une réparation* ». Nous pouvons ajouter des informations complémentaires afin de mieux expliquer la situation, par exemple « *la réparation concerne un équipement électrique* ». Ainsi, le type d'équipement peut être considéré comme une information interne ou une description (partielle) de la réparation elle-même, et peut être mise « à l'intérieur » du nœud représentant la réparation. Après avoir ajouté cette information dans le nœud « réparation », le deuxième graphe conceptuel de la Figure II.4 est obtenu ; il s'agit maintenant d'un graphe conceptuel emboîté représentant un plus haut niveau d'abstraction.

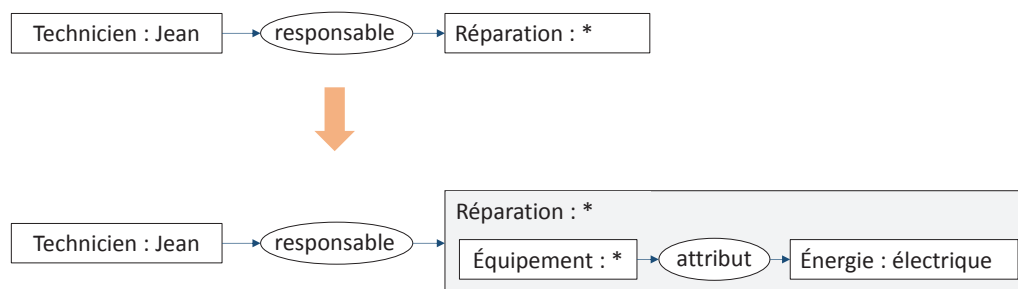


Figure II.4. Illustration d'un graphe conceptuel simple et d'un graphe conceptuel emboîté

## II.2.6 INTERET DES GCs DANS NOTRE APPROCHE REX-ECD

Le principal intérêt de ce formalisme de représentation de connaissances est sa capacité d'expressivité et de raisonnement avec les graphes, aspect clé pour la structuration sémantique de la base de REX-ECD en vue d'une exploitation future par les utilisateurs. Il s'agit d'un formalisme qui permet une représentation unifiée des différents types de connaissances (e.g. connaissances du domaine, expériences passées, règles extraites, relations entre règles, requêtes de l'utilisateur...) que nous allons considérer dans notre approche conceptuelle REX-ECD comme support de prise de décision.

## II.3 DESCRIPTION DE LA DEMARCHE GLOBALE REx-ECD

Le retour d'expérience peut s'appliquer à différents types d'activités de l'entreprise, et peut en particulier permettre d'améliorer les performances des processus (Chebel-Morello, 2008). Nous nous intéressons principalement ici aux expériences techniques ayant eu lieu sur les processus de production, généralement enregistrées dans des bases de données techniques. C'est à partir de ces bases de données disponibles que nous envisageons la formalisation et la structuration des expériences déjà recueillies à l'aide d'un formalisme de représentation, puis l'extraction de nouvelles connaissances sous la forme de règles, afin de constituer un référentiel d'entreprise ou un référentiel métier (support à une connaissance décisionnelle) dans un contexte particulier.

Un avantage important du processus d'extraction de connaissances est que les données nécessaires pour l'analyse peuvent être collectées au cours des opérations normales du processus de production ; il n'est donc pas indispensable de définir un processus particulier dédié à la collecte des données (Harding et al., 2005).

Dans ce contexte, l'extraction de connaissances procédurales sous la forme de règles peut être utile pour les deux raisons précédemment abordées : *i)* les motifs extraits sont facilement compréhensibles pour un utilisateur non-expert, et *ii)* la charge de travail de l'utilisateur humain pendant l'extraction reste limitée, sauf lors de l'évaluation et de la validation des résultats. Ainsi, afin de générer ces règles, une information brute ou semi-structurée sera d'abord sélectionnée à partir des bases de données considérées. Ensuite, cette information devra être « nettoyée » afin de la rendre exploitable, et finalement, ces données seront traitées pour en tirer de nouvelles connaissances associées.

L'ECD est ainsi considérée comme complémentaire à la démarche de REx, qui est elle aussi facilement intégrable aux processus opérationnels existants. L'interaction entre ces deux démarches implique un échange important d'informations (Rakoto, 2004). A notre avis, le couplage entre ces deux approches (REx et ECD) n'a pas été suffisamment étudié dans la littérature, malgré leur complémentarité. C'est donc le principal objectif de notre travail.

En outre, nous proposons de distinguer dans notre approche deux niveaux différents dans la base de REx-ECD : une « **base d'expériences** » concernant les connaissances spécifiques qui caractérisent les expériences passées, et une « **base de règles** » contenant une connaissance plus générale issue de l'analyse des informations disponibles, toutes deux ayant en commun une connaissance ontologique. Dans ce cadre conceptuel, les principales phases du processus de REx proposé, intégrant les phases du processus d'ECD, sont schématisées sur la Figure II.5 et sont décrites plus en détail dans les parties suivantes (II.3.1, II.3.2 et II.3.3), de même que les outils utilisés et les flux d'information envisagés entre les différentes étapes.

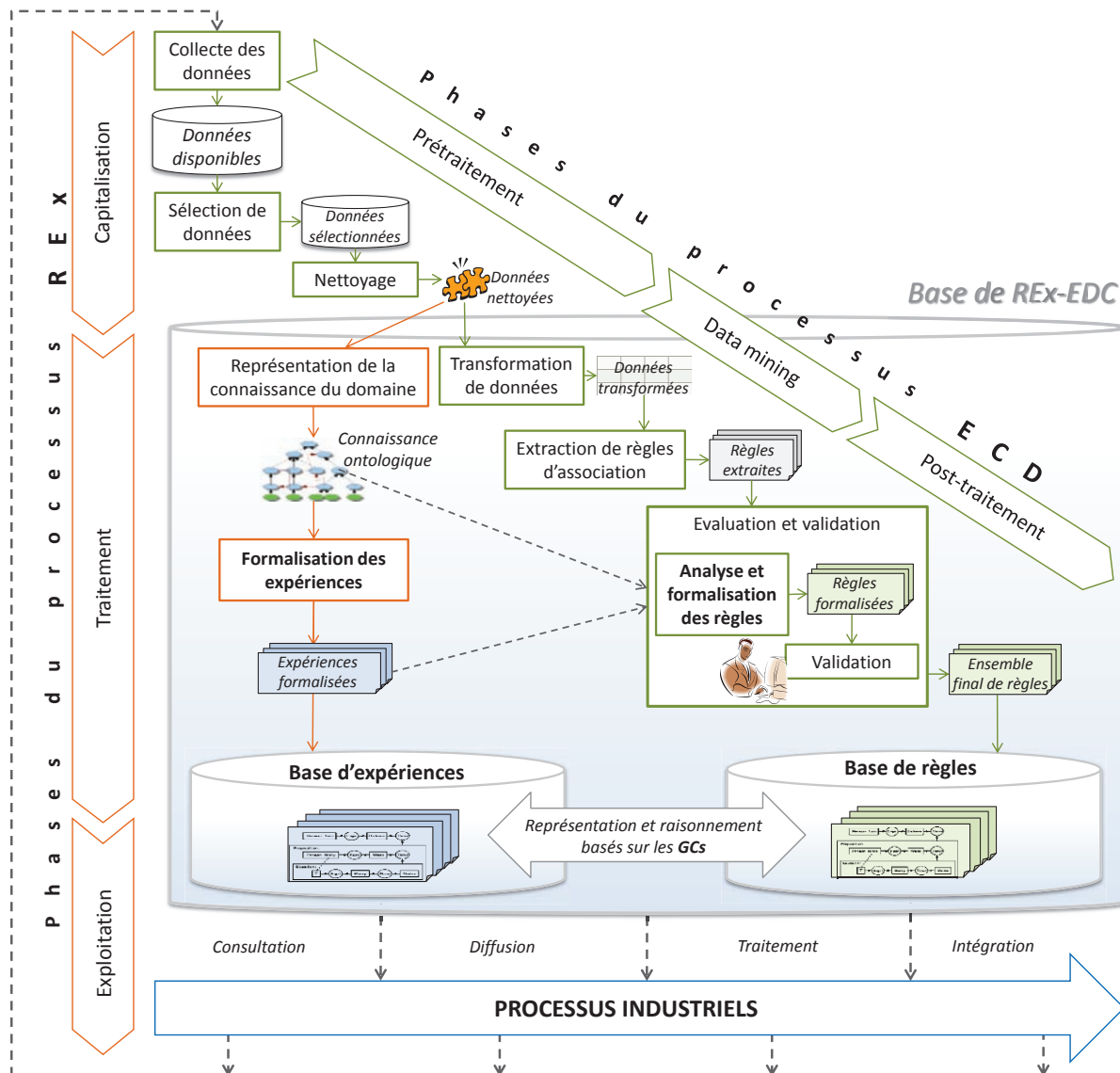


Figure II.5. Principales étapes de la démarche conceptuelle REx-ECD

### II.3.1 PRETRAITEMENT DE L'ECD ASSOCIE A LA PHASE DE CAPITALISATION DU REX

Les premières étapes de notre approche (en haut à gauche sur la Figure II.5) concernent l'intersection entre la phase de capitalisation du REx proposé et la phase de prétraitement de l'ECD.

Tout d'abord, nous devons comprendre dans cette phase le domaine d'application, identifier les objectifs et vérifier les besoins du système afin de définir une cible et de *sélectionner* l'information (transactions, attributs ou variables) pertinente pour être traitée, puisque toutes les données disponibles et accessibles ne sont certainement pas utiles dans leur intégralité (Zighed & Rakotomalala, 2002). De ce fait, il semble intéressant ici d'envisager une première interaction avec l'expert du domaine d'application, portant surtout sur la définition des objectifs du processus d'extraction, pour cibler l'espace des données qui va être exploré selon le point de vue de l'expert.

Une étape de *nettoyage des données* est ensuite nécessaire afin de détecter et de corriger les inexactitudes et/ou erreurs dans les données avant les phases d'analyse et de traitement. Cette partie vise à améliorer la qualité des données, car ces données du « monde réel » sont généralement incomplètes (valeurs manquantes, attributs d'intérêt manquants, etc.), imprécises (contenant des erreurs ou des doublons), ou incohérentes (contenant des anomalies). Cette étape implique ainsi l'utilisation de techniques telles que le nettoyage de données, la réduction de données ou la discrétisation (Köksal et al., 2011). Cette partie du processus demande un temps d'exécution important, mais son déroulement est déterminant pour la suite du processus d'ECD proposé. Une mauvaise qualité des données pourrait en effet entraîner une mauvaise qualité des résultats de l'exploitation (Han & Kamber, 2001), d'où l'importance de ces premières étapes.

Nous sommes particulièrement intéressés par les données transactionnelles dans les entreprises, i.e. par les bases de données composées d'un ensemble d'enregistrements décrits par un identifiant (ID) et un ensemble de valeurs d'attributs ou *items*. Ainsi par exemple, si pour certains attributs des valeurs sont manquantes ou incohérentes, cette étape de nettoyage vise à les calculer ou à les déterminer en tenant compte du type d'attribut (quantitatif ou qualitatif) et de l'ensemble des données présentes. Nous présentons dans la Figure II.6 un exemple simplifié pour illustrer les premières étapes du prétraitement dans la démarche REx-ECD consistant à sélectionner, inférer et/ou corriger des données.

Dans les parties suivantes, nous présentons en détail les phases de traitement et *data mining* (II.3.2) et d'exploitation et post-traitement (II.3.3) de la démarche REx-ECD.

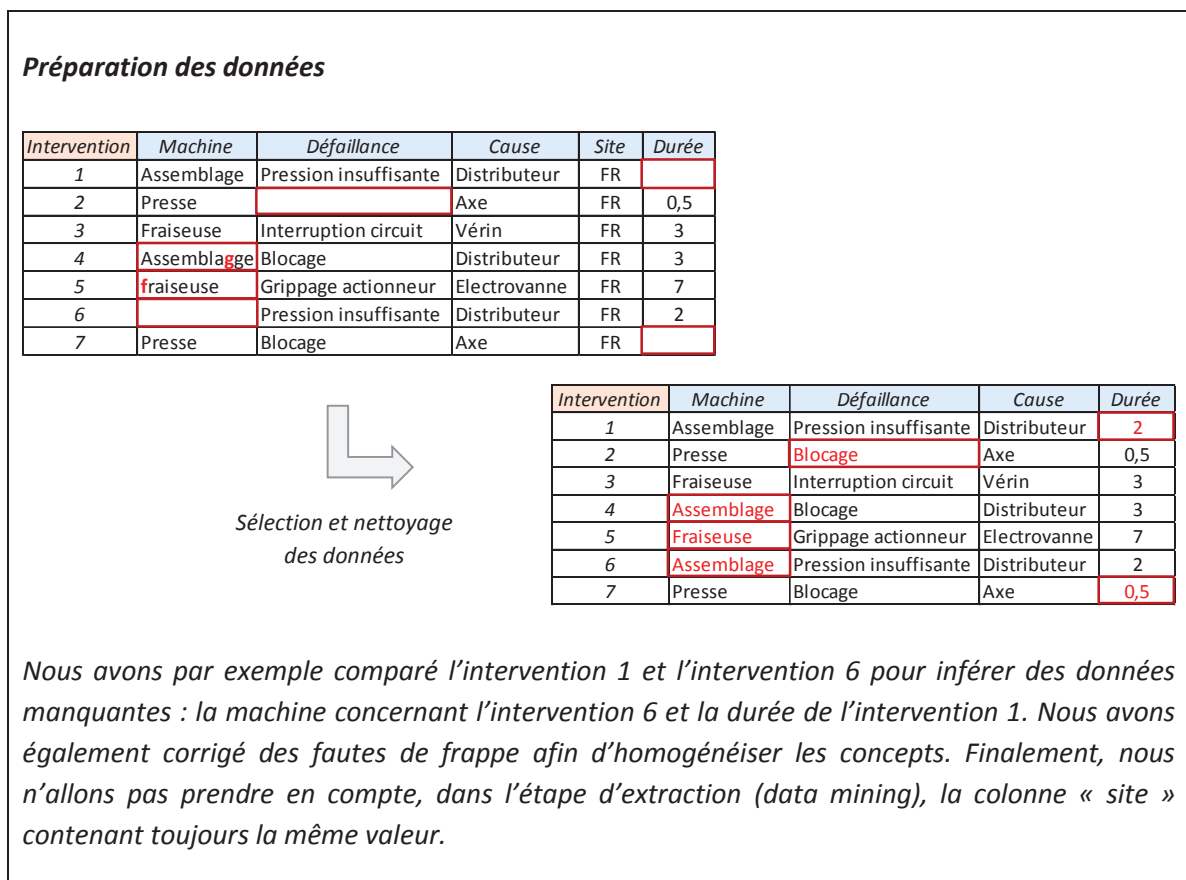


Figure II.6. Illustration de la préparation des données

### II.3.2 TECHNIQUES DE DATA MINING ASSOCIEES A LA PHASE DE TRAITEMENT DU REx

Une fois que les données ont été sélectionnées et nettoyées, la phase de traitement de l'extension du système REx porte sur deux défis principaux :

- La structuration d'une « **base d'expériences** » (chemin de gauche de la Figure II.5) en vue d'une réutilisation future. Une expérience étant considérée dans notre contexte comme une instance singulière de la connaissance, celle-ci pourrait sembler partiellement implicite dans les bases de données. Toutefois, une meilleure formalisation des informations caractérisant les expériences, ici à l'aide des GCs construits à partir d'un vocabulaire du domaine, rend ces connaissances plus explicites et exploitables. Ces connaissances explicitées pourraient tout d'abord fournir un support important dans la résolution de problèmes et dans la prise de décisions (voir section II.3.3.1). D'autre part, ces expériences, ainsi que la connaissance formalisée du domaine, vont aussi aider l'évaluation et la validation des connaissances extraites dans le processus d'ECD proposé (cette dernière partie sera abordée plus en détail dans le Chapitre III).
- La génération d'une « **base de règles** » (chemin de droite de la Figure II.5) à partir de l'analyse des informations qui caractérisent les expériences, en vue de diffuser et d'intégrer de nouvelles connaissances plus génériques sous forme de règles dans le processus industriel. A cet effet, nous allons considérer un algorithme d'extraction de règles d'association pour la découverte des différents types de règles, puis une méthodologie pour faciliter leur interprétation et leur évaluation. Finalement, le mécanisme de raisonnement privilégié dans les GCs (Mugnier, 2000) sera considéré afin de fournir à l'utilisateur final les moyens de mieux explorer et filtrer les règles extraites selon ses intérêts et/ou besoins du moment.

#### II.3.2.1 Structuration de la base d'expériences à l'aide des graphes conceptuels

Dans cette partie, nous allons présenter les principales étapes développées pour aboutir à une base d'expériences formalisées (Figure II.7).

Dans le contexte du REx, une expérience concerne « *l'ensemble des informations générées, permettant de caractériser ce qui s'est passé depuis l'apparition de l'événement* » (Bertin, 2012) ; plus particulièrement, elle concerne la description de l'événement déclencheur et l'analyse effectuée pour aboutir à des solutions. De ce fait, pour mieux représenter une expérience  $E_i$ , un vecteur composé de trois champs est défini :  $E_i = \langle C_i, A_i, S_i \rangle$ , où  $C_i$ ,  $A_i$  et  $S_i$  représentent respectivement le contexte dans lequel l'événement est survenu, l'analyse menée et la solution mise en œuvre. Ces trois lots d'information sont décrits par un ou plusieurs attributs. Une base d'expériences correspond donc à un ensemble d'expériences :  $B_{Exp} = \{E_i, i \in \{1 \dots n\}\}$ .

Afin de formaliser les connaissances attachées aux expériences à l'aide des GCs, il est proposé dans (Chein & Mugnier, 1992) de définir d'abord le « support » ( $S$ ), qui facilite le partage, la recherche, la gestion et l'exploitation des connaissances.



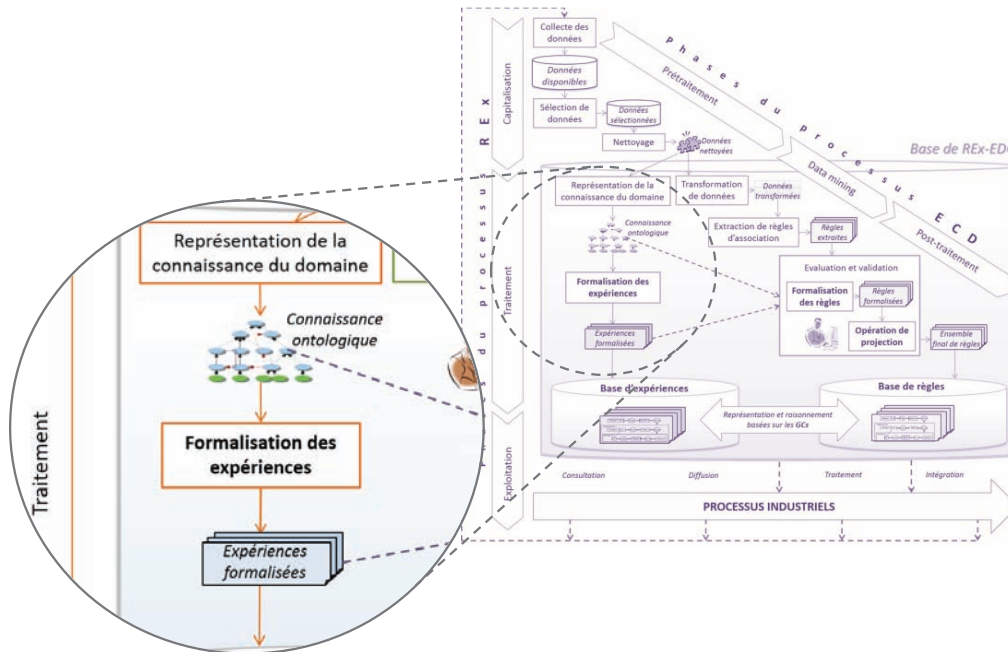


Figure II.7. Principales étapes pour la structuration de la base d'expériences

### Définition et représentation de la connaissance ontologique

Les composants d'une ontologie au niveau conceptuel devraient permettre aux organisations de formaliser les différents types de connaissances utilisées au niveau opérationnel. En d'autres termes, ce sont les hiérarchies des types de concepts et de relations qui vont déterminer l'expressivité des connaissances procédurales dans la base de connaissances (ici la base de REx-ECD). Plus particulièrement, cette connaissance ontologique est essentiellement utilisée ici pour faciliter la formalisation des expériences, et le sera plus tard pour la représentation, l'exploitation et la réutilisation des connaissances sous la forme de règles.

### Formalisation des connaissances attachées aux expériences

Au niveau opérationnel, les GCs vont représenter la connaissance confinée dans chaque expérience, i.e. les informations caractérisant l'expérience. Ils montrent notamment quels axiomes sémantiques sont requis pour utiliser la connaissance ontologique de manière opérationnelle (Fürst et al., 2003), et permettent de satisfaire la contrainte d'intégrer les connaissances d'une manière qui facilite le partage et la réutilisation.

A partir d'un support (*S*) prédéfini pour la démarche REx, nous avons construit, à l'aide des GCs, un modèle générique élémentaire pour représenter les expériences (Figure II.8), i.e. une représentation structurée d'un événement survenu dans un contexte auquel sont associés une analyse et des actions de résolution menées. Il s'agit d'un modèle générique qui devrait être adapté, et notamment enrichi, en fonction des caractéristiques, des contraintes et des restrictions du contexte d'application.

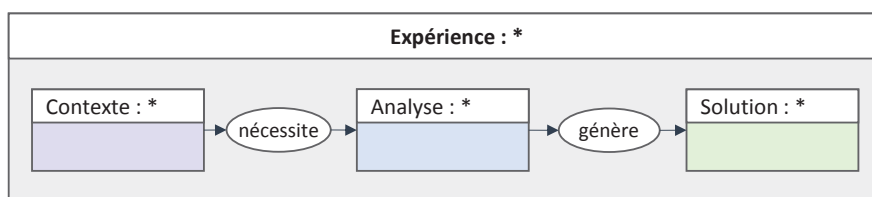


Figure II.8. Modèle générique pour représenter une expérience

Finalement, une fois que les expériences ont été formalisées, l'étape suivante est la mise en mémoire (stockage) dans la base d'expériences.

### II.3.2.2 Génération des règles d'association à partir d'une technique de data mining

Au sein du processus d'extraction de nouvelles connaissances, Silberschatz et Tuzhilin (Silberschatz & Tuzhilin, 1996a) proposent des manières d'intégrer l'utilisateur dans l'analyse, notamment au travers de trois types de processus d'extraction : automatique, semi-automatique ou manuel.

Dans notre démarche, le processus d'extraction de règles d'association concerne, dans une certaine mesure, l'intervention de l'utilisateur humain afin de guider le processus et de valider les résultats ; il s'agit ainsi d'un processus semi-automatique dans lequel l'implication de l'utilisateur tout au long du développement de l'ECD devient essentielle afin que le système réponde à ses besoins et/ou attentes.

Toutefois, une implantation convenable des techniques de *data mining* dans les entreprises n'est pas un processus facile à mettre en place. En effet, il existe différentes limites ou difficultés qui peuvent être rencontrées lors de la mise en œuvre du processus d'ECD (Wang, 2007), liées par exemple à la disponibilité des ensembles de données, à la propreté de ces données (un prétraitement étant toujours nécessaire), à la pertinence du contenu de l'information, aux relations entre les données, à la confidentialité dans les entreprises (qui rend parfois difficile la collecte de données et l'interprétation des résultats), à la difficulté d'interpréter les résultats obtenus (puisqu'ils sont spécifiques à un domaine donné, et nécessitent donc l'intervention d'experts du domaine d'application), à la sélection des outils et des logiciels, entre autres (Kasravi, 1997).

Sur le schéma de droite de la Figure II.5 ont été présentées les principales phases du processus d'ECD suggéré, i.e. le prétraitement, le *data mining* et le post-traitement. Nous nous concentrons dans cette partie sur deux étapes particulières du processus : la transformation des données nettoyées et l'extraction des règles d'association (Figure II.9).

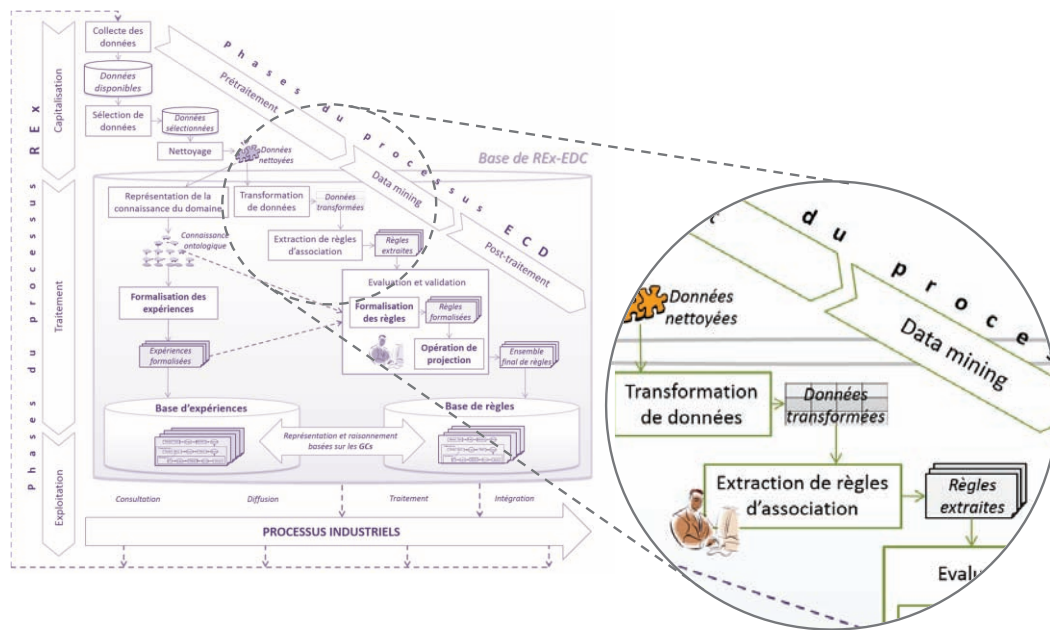


Figure II.9. Un prétraitement avant le data mining

### Étape de transformation des données

Nous avons déjà présenté les premières étapes concernant la préparation des données, i.e. la sélection et le nettoyage des données.

Nous abordons maintenant l'étape de transformation des données, qui vise à organiser les données dans un format adéquat pour l'application d'un algorithme de *data mining*, permettant la découverte des règles d'association. La transformation consiste à formater les données (souvent sous forme tabulaire) en vue de l'opération d'extraction.

Pour l'application d'un algorithme d'extraction de règles d'association, nous devons construire un « contexte formel » à partir des transactions disponibles (expériences recueillies), tout en tenant compte de la connaissance du domaine contenue dans le support ( $S$ ). Ce contexte formel est défini par un triplet  $B^* = (T, I, R)$ , où  $B^*$  représente la base de données réduite et/ou nettoyée pour la transformation,  $T$  est un ensemble de transactions,  $I$  est un ensemble d'*items* (ici représentés par les types concepts définis dans le support) et  $R \subseteq T * I$  est une relation binaire entre  $T$  et  $I$ . Nous présentons dans la Figure II.10 une illustration de l'étape de transformation des données dans la démarche REx-ECD.

### Mise en forme tabulaire des données

Après l'étape de nettoyage, la mise en forme tabulaire des données d'entrée d'un algorithme d'extraction de règles est nécessaire. Pour illustrer cette étape, nous construisons ici, à partir du format de données initiales, un contexte formel avec 6 items et 10 transactions.

Le contexte formel créé  $B^* = (T, I, R)$  est présenté sur la partie de droite de l'illustration ci-dessous. Les colonnes représentent les items  $I$  (de A à F) et les lignes représentent les transactions  $T$  (de T1 à T10). Nous représentons également les attributs présents et absents contenus dans  $I$  (« x » correspond à un attribut présent et « o » représente un attribut absent).

Transaction	Liste de items										
T1	A	B	E								
T2	B	D									
T3	B	C									
T4	A	B	D								
T5	A	C									
T6	B	C									
T7	A	C									
T8	A	B	C	E							
T9	A	B	C								
T10	A	B	E	F							

Transformation vers le  
contexte formel

R	A	B	C	D	E	F
T1	x	x	o	o	x	o
T2	o	x	o	x	o	o
T3	o	x	x	o	o	o
T4	x	x	o	x	o	o
T5	x	o	x	o	o	o
T6	o	x	x	o	o	o
T7	x	o	x	o	o	o
T8	x	x	x	o	x	o
T9	x	x	x	o	o	o
T10	x	x	o	o	x	x

Figure II.10. Illustration de la transformation des données

### Étape de découverte des *itemsets* (ensembles d'*items*) fréquents et génération des règles d'association

Dans la phase de *data mining* du processus d'ECD, nous nous intéressons à la découverte des relations existantes dans les données, afin de mieux comprendre le système étudié. Ces relations concernent le comptage de cooccurrences et sont classiquement nommées « règles d'association ».

L'extraction de règles d'association concerne ainsi la découverte des relations pertinentes, des corrélations, des motifs fréquents ou des associations entre des ensembles d'*items* sélectionnés a priori. L'objectif ici est de dériver des règles d'association (Agrawal et al., 1993) à partir des informations décrivant les expériences passées, ces règles reliant les *items* ou valeurs d'attributs fréquents du domaine modélisé.

Rappelons qu'en dépit du vocabulaire utilisé dans le domaine de l'extraction des règles d'association, de telles règles ne désignent pas des implications ou liens de causalité (i.e. « cause-effet ») entre deux ensembles d'*items*  $X$  et  $Y$  ; elles soulignent plutôt des corrélations, ou la fréquence d'apparition d'un ensemble d'*items* donné dans une transaction, qui inclurait la présence d'un autre ensemble d'*items* dans la même transaction (Agrawal et al., 1993). Par exemple, la règle d'association  $X \rightarrow Y$

nous indique que lorsqu'une transaction contient l'ensemble d'*items*  $X$ , alors l'ensemble d'*items*  $Y$  est souvent rencontré, sans conclure qu'il existe une relation de cause à effet entre les deux ensembles d'*items* corrélés.

De nombreux algorithmes existent pour découvrir ces règles d'association (section I.5.3.4) contenues dans une base de données transactionnelle, le plus connu étant l'algorithme *Apriori*.

En raison de sa popularité et de ses bonnes performances, nous avons choisi l'algorithme classique *Apriori* pour l'extraction de règles d'association dans nos exemples d'application (Chapitre IV). Contrairement à certains algorithmes, *Apriori* n'impose pas de contraintes sur la forme des règles extraites. Au contraire, il trouve plusieurs types de règles d'association entre les ensembles d'*items* (*itemsets*) fréquents. L'algorithme construit ces *itemsets* fréquents en ajoutant seulement des *items* fréquents aux *itemsets* fréquents, et en supprimant tous les *itemsets* non fréquents (Chiang & Miller, 2008). Ceci reste cohérent avec notre objectif d'extraire différents types de règles, pour une évaluation et validation ultérieure des résultats par l'utilisateur expert du domaine.

Pour aborder le problème général d'extraction de règles d'association, dans lequel  $m$  *items* conduisent potentiellement à  $2^m$  *itemsets*, l'algorithme *Apriori* (Agrawal & Srikant, 1994) utilise une procédure d'estimation pour déterminer les *itemsets* dont la fréquence devrait être évaluée à chaque itération en fonction d'un seuil minimal de support (*minsup*). En effet, après l'élagage (lignes 13 à 18 dans l'Algorithme I.1), un contrôle est effectué afin que lors d'une itération, tous les éléments mesurés ne soient pas stockés. Ainsi, un *itemset*  $X$  candidat de taille  $k$  est « fréquent », si et seulement si chaque sous-ensemble de  $X$ , est aussi fréquent. Cette considération permet une réduction significative de l'espace de recherche, et permet l'extraction des règles dans un temps de calcul raisonnable (Nahar et al., 2013).

Un autre problème de taille exponentielle est celui de la génération des règles à partir des *itemsets* fréquents. Le temps d'exécution ici est toutefois moins important que dans l'étape précédente : dans ce cas, pour un  $k$ -*itemset* (*itemset* de taille  $k$ ), le nombre de règles d'association possible est  $2^k - 2$ . Dans ce contexte, identifier les règles intéressantes devient très compliqué, de par le grand nombre de règles découvertes par les algorithmes d'extraction, notamment pour les ensembles de transactions dont les attributs sont fortement corrélés. De ce fait, une autre mesure permettant de pré-filtrer les règles extraites, est intégrée dans l'algorithme *Apriori* : le *minconf* (seuil minimal de confiance) (voir Chapitre I).

Dans notre approche, ces deux mesures procurent une première manière d'évaluer les règles dans le processus. Elles sont prises en compte dans les deux étapes fondamentales de l'algorithme *Apriori* (Algorithme I.1. et Algorithme I.2) pour l'extraction des règles d'association (Agrawal & Srikant, 1994) :

- i) la découverte des *itemsets* satisfaisant le **support minimum** spécifié par l'utilisateur à partir d'une base de données (recherche d'*itemsets* fréquents), et
- ii) la génération des règles d'association satisfaisant la **confiance minimum** requise par l'utilisateur à partir des *itemsets* fréquents trouvés.

Le nombre de règles, et donc leur généricité, sera en effet piloté en définissant des seuils prédéfinis *minsup* et *minconf*, désignant le support minimum et la confiance minimum d'une règle (Agrawal et al., 1993). Une règle est ainsi considérée comme « fréquente » si son support est supérieur ou égal au seuil *minsup* prédéfini, i.e.  $Sup(X \rightarrow Y) \geq minsup$  ; et « forte » ou « robuste » si la confiance est supérieure ou égale au seuil *minconf* prédéfini, i.e.  $Conf(X \rightarrow Y) \geq minconf$ .

Avant de présenter la façon avec laquelle fonctionne l'algorithme *Apriori* (Algorithme I.1. et l'Algorithme I.2), nous présentons dans la Figure II.11 quelques résultats obtenus à partir de l'exemple détaillé dans la Figure II.10 lorsque le *minsup* et le *minconf* varient.

#### Résultats obtenus lorsque le *minsup* et le *minconf* varient

A partir du contexte formel construit (partie de droite de la Figure II.10), nous allons extraire des règles d'association en tenant compte des seuils prédéfinis. Nous présentons dans le tableau ci-dessous quelques résultats obtenus en fonction du *minsup* et du *minconf*. Nous pouvons constater que le nombre d'itemsets fréquents dépend directement du *minsup* et que le nombre de règles dépend directement du *minconf* et du nombre d'itemsets fréquents.

Minsup	20%	20%	30%	30%	40%	40%
Minconf	50%	100%	50%	100%	50%	100%
Itemsets fréquents	13	13	10	10	6	6
Règles extraites	15	6	12	5	6	0

Considérons par exemple un *minsup* = 30% et un *minconf* = 50%. Ces valeurs ont conduit à l'extraction de 10 itemsets fréquents et de 12 règles d'association.

Figure II.11. Illustration des résultats obtenus en fonction du *minsup* et du *minconf*

Afin d'illustrer le processus d'extraction des règles d'association, la première étape de l'algorithme concernant la génération des *itemsets* fréquents est d'abord présentée (Figure II.12), la deuxième étape concernant la découverte des règles étant illustrée dans la Figure II.13.

Finalement, une phase de post-traitement de ces règles extraites, qui présente à notre avis un intérêt spécifique, est nécessaire pour interpréter, évaluer et filtrer les règles extraites d'un point de vue plus subjectif. Étant donné que l'utilisation seule des deux mesures présentées (support et confiance) ne suffit à notre avis pas à garantir la qualité des règles détectées, dans l'étape suivante, l'utilité et l'intérêt de la règle sont décidés par l'utilisateur du domaine industriel.

Nous présentons dans la partie suivante cette dernière phase de la démarche d'ECD ainsi que la phase d'exploitation du REX (section II.3.3).

**Première étape : génération des itemsets fréquents**

En choisissant un minsup de 30%, nous présentons ci-dessous la première partie de l'algorithme Apriori afin de générer les itemsets fréquents satisfaisant le support minimum prédéfini.

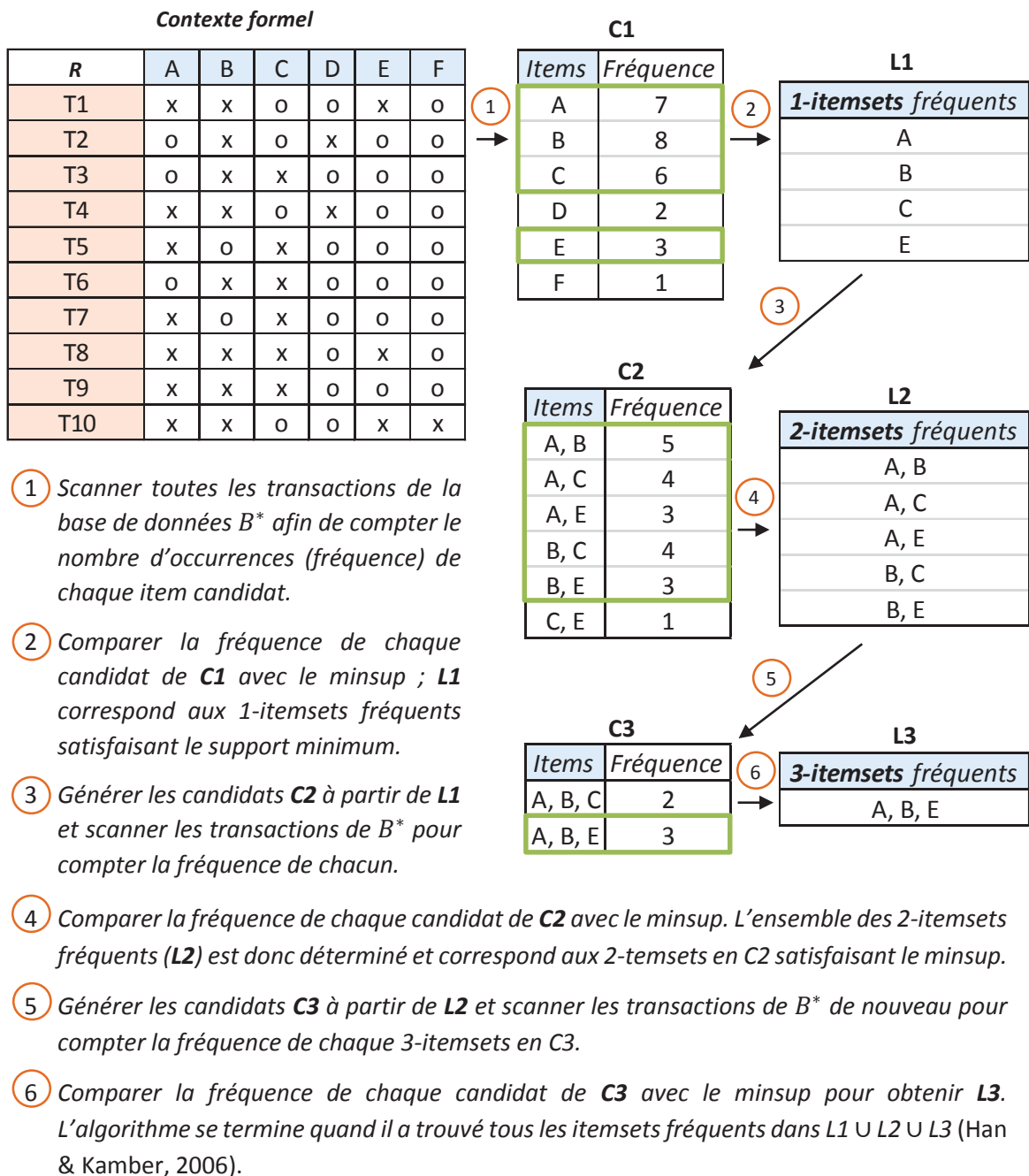


Figure II.12. Illustration de la génération des itemsets candidats et des itemsets fréquents



**Deuxième étape : génération des règles d'association à partir des itemsets fréquents trouvés**

A partir de l'ensemble  $(L1 \cup L2 \cup L3)$  de tous les itemsets fréquents trouvés dans le contexte d'extraction  $B^*$ , les règles d'association sont générées. Nous présentons ici la deuxième partie l'algorithme Apriori en tenant compte d'un minconf de 50%.

L'ensemble des règles d'association extraites est présenté dans le tableau ci-dessous :

Règle #	Hypothèse	Conclusion	Sup (%)	Conf (%)
R1	E	B	30	100
R2	E	A	30	100
R3	B	C	40	50
R4	C	B	40	67
R5	A	C	40	57
R6	C	A	40	67
R7	A	B	50	71
R8	B	A	50	62
R9	E, A	B	30	100
R10	E, B	A	30	100
R11	A, B	E	30	60
R12	E	A, B	30	100

Pour chaque itemset fréquent  $l_k$ , tous les sous-ensembles  $a_{m-1}$  de  $l_k$  sont déterminés. Prenons un 2-itemset fréquent de la liste, contenant par exemple les items  $\{B, E\}$ . La règle d'association  $B \rightarrow E$  n'a pas été générée car la valeur de  $Sup(B \rightarrow E)/Sup(B)$  est inférieure au seuil minconf (50%) :

- $Support(B \rightarrow E) = 3/10 * 100\% = 30\%$
- $Confiance(B \rightarrow E) = Support(B \rightarrow E)/Support(B) = 30/80 * 100\% = 37,5\%$

Par contre, la règle inverse  $(E \rightarrow B)$ , règle R1 dans le tableau ci-dessus, a été générée avec une confiance de 100% :

- $Support(E \rightarrow B) = 3/10 * 100\% = 30\%$
- $Confiance(E \rightarrow B) = Support(E \rightarrow B)/Support(E) = 30/30 * 100\% = 100\%$

**Figure II.13. Illustration de la génération des règles d'association**



### II.3.3 POST-TRAITEMENT DE L'ECD ASSOCIE A LA PHASE D'EXPLOITATION DU REX

La dernière phase du processus de REx, i.e. la phase d'exploitation, concerne la valorisation et la mise en action par les utilisateurs du domaine des connaissances spécifiques (i.e. expériences) et des connaissances plus générales (i.e. règles) contenues dans la base de REx-ECD.

D'un point de vue général, la base de REx-ECD proposée est fournie principalement pour la consultation, la diffusion, le traitement ou l'intégration de connaissances traitées issues des processus industriels. Nous présentons dans la Figure II.14 les différentes utilisations potentielles de cette base de connaissances :

- tout d'abord, la base de REx-ECD peut servir de source d'information générale par un filtrage de l'information ;
- une autre utilisation, plus originale, consiste à utiliser celle-ci comme un support d'aide à la décision dans les processus industriels ;
- finalement, une autre exploitation de la base concerne l'utilisation des connaissances extraites pour améliorer les processus. Cette dernière utilisation implique notamment une validation a priori des connaissances par l'utilisateur du domaine et leur mise en action, selon le contexte et les besoins du domaine d'application.

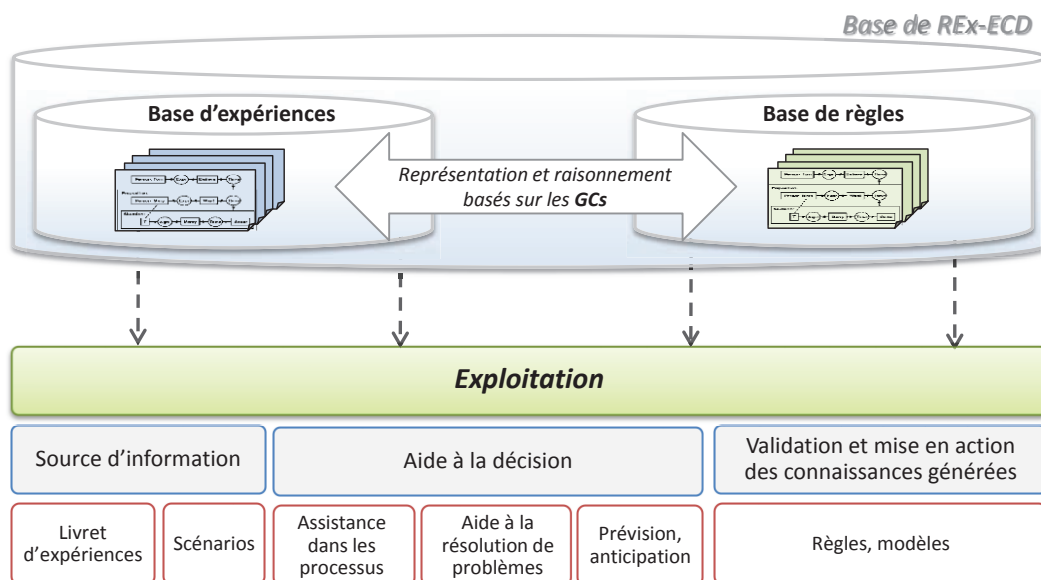


Figure II.14. Exploitation possible de la base de REx-ECD (adaptée de (Béler, 2008))

Nous nous intéressons aux deux dernières utilisations abordées, i.e. aide à la décision, et validation et mise en œuvre des connaissances générées. Plus particulièrement, nous allons présenter une première exploitation possible de la base d'expériences formalisées comme un point essentiel d'aide à la décision dans la résolution de futurs problèmes (II.3.3.1). Nous allons ensuite aborder dans cette

phase d'exploitation du REx notre principale contribution au processus d'ECD, notamment lors du post-traitement des connaissances extraites, qui implique l'interprétation, la visualisation et la validation des règles générées pour une future réutilisation (II.3.3.2).

### **II.3.3.1 Une première exploitation possible de la base d'expériences formalisées**

Dans une première démarche REx décrite dans (Potes Ruiz et al., 2013), nous avons proposé une méthodologie basée sur le Raisonnement à Partir de Cas (RàPC) dans laquelle les GCs sont utilisés pour modéliser les expériences dans le cadre d'un RàPC collaboratif, en prenant en considération l'avis des experts<sup>4</sup>. Le principal objectif de cette méthodologie basée sur le retour d'expérience classique porte uniquement sur la réutilisation des expériences passées pour assister les utilisateurs finaux dans la résolution d'un nouveau problème. Bien que cette première expérience ait des motivations un peu différentes de celles de ce travail de thèse qui a une plus vaste portée, nous pensons intéressant d'en rappeler les aspects principaux dans la suite visant une future exploitation de la base d'expériences formalisées contenue dans la base de REx-ECD (Figure II.14).

Dans cette première étude, nous avons proposé l'utilisation étendue d'un RàPC en profitant des mécanismes de raisonnement des GCs afin d'identifier les expériences passées proches du cas à traiter. Le principal objectif était de retrouver dans la base d'expériences les expériences les plus proches du nouveau problème (en tenant compte de leur contexte et de leur analyse), et de mettre en œuvre une solution au problème à résoudre (après son adaptation éventuelle).

D'une manière générale, le cycle du RàPC collaboratif proposé dans (Potes Ruiz et al., 2013) est illustrées sur la Figure II.15. Soulignons que dans un contexte de RàPC, l'expérience s'exprime sous la forme de cas.

Les étapes du RàPC collaboratif sont :

- 1) *Elaboration du cas cible* (désigné aussi comme « cas courant » ou « nouveau cas ») qui implique la construction d'un nouveau cas en prenant comme base les connaissances du domaine (via le support ontologique) et le modèle générique de représentation d'une expérience (Figure II.8).
- 2) *Recherche des cas sources* (désignés aussi par « cas passés » ou « cas de la base ») proches ou similaires. Une mesure de similarité sémantique a été ainsi définie pour sélectionner, parmi tous les cas sources de la base, les cas les plus proches. Cette mesure est nécessaire afin de représenter numériquement les informations de ressemblance entre les différentes expériences (Ruet, 2002).
- 3) *Agrégation d'avis des experts* qui consiste à agréger et exploiter ces cas identifiés (Potes Ruiz et al., 2012). L'agrégation d'avis d'experts devient donc l'étape qui détermine comment les

---

<sup>4</sup> Ces travaux ont fait l'objet de mon stage de Master Recherche au sein du Laboratoire Génie de Production.

solutions de l'ensemble des cas sélectionnés seront combinées dans la solution finale (Leake et Sooriamurthi, 2002).

- 4) *Adaptation*, qui consiste à construire la solution au problème du cas cible à partir de la solution trouvée sur les cas sources.
- 5) *Révision*, consistant à évaluer la pertinence de la solution suggérée pour le cas cible et à la modifier, si nécessaire, avant de la valider. Le cas révisé et validé devient une nouvelle expérience qui doit être capitalisée dans la base d'expériences.
- 6) *Stockage*, dernière étape du mécanisme, dans laquelle la nouvelle expérience est mémorisée. La nouvelle expérience obtenue est ainsi incorporée à la base d'expériences pour une réutilisation future.

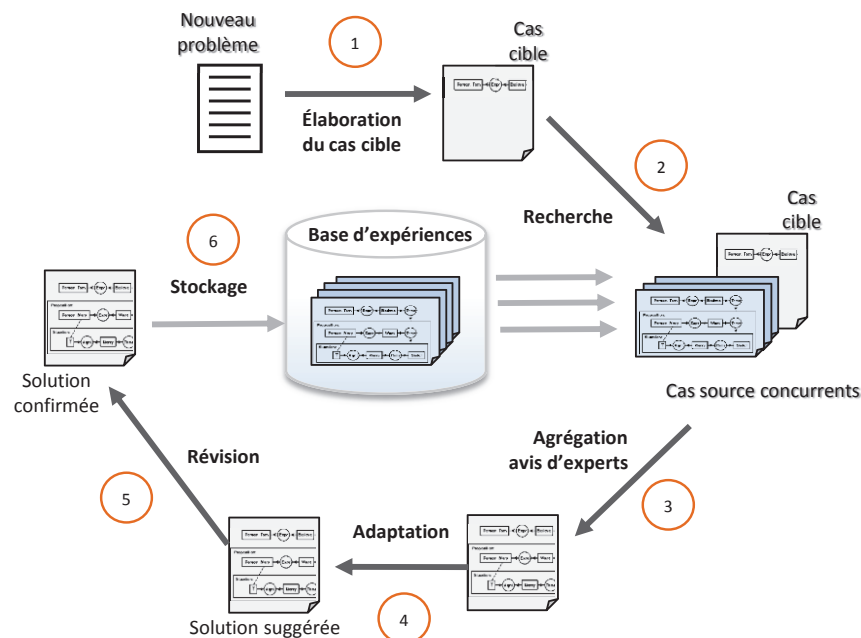


Figure II.15. Une réutilisation possible de la base d'expériences formalisées

En dépit de l'intérêt initial manifesté sur ce sujet, nous avons décidé dans ce travail de thèse d'élargir la démarche classique de REx qui prend souvent en considération les techniques du RàPC pour adapter la connaissance contenue dans les expériences passées. La principale motivation de ce travail de thèse étant de générer des nouvelles connaissances plus générales que les expériences, nous allons nous concentrer par la suite sur l'analyse et l'exploitation possible des connaissances extraites (qui seront stockées dans la base de règles - Figure II.14) afin d'aider la prise de décision stratégique et d'améliorer un référentiel de connaissances globales des entreprises, ce qui ne relève pas des approches classiques de retour d'expérience compte tenu des volumes importants d'informations collectés quotidiennement à partir des processus industriels.

### II.3.3.2 Post-traitement des règles extraites pour une future réutilisation

La technique d'extraction de règles d'association en utilisant l'algorithme *Apriori* (fondé sur le support et la confiance) présente certains problèmes, tels que la tendance à découvrir un grand nombre de règles d'association ou le fait que toutes les règles découvertes ne sont pas pertinentes ou intéressantes pour les experts du domaine d'application.

Une évaluation et une validation des résultats générés par l'algorithme d'extraction a mis en avant la nécessité d'une phase de post-mining dans le processus général d'ECD, notamment afin d'améliorer la qualité de la connaissance extraite.

Cette dernière phase du processus d'ECD porte principalement sur l'interprétation, la visualisation des résultats obtenus, l'évaluation et la validation du point de vue de l'utilisateur expert du domaine. Celui-ci doit en effet pouvoir tester différentes hypothèses ou préférences afin de trier les résultats suivant ses besoins/attentes.

Nous insistons ici sur la contribution de l'utilisateur expert du domaine qui a, en effet, la responsabilité d'évaluer et de valider les nouvelles connaissances extraites, à l'aide des techniques de visualisation choisies qui l'assisteront dans le processus. Cette dernière problématique fait l'objet du Chapitre III.

## II.4 BILAN

La démarche REx-ECD proposée, ainsi que ses outils de mise en œuvre, a été conçue principalement pour supporter la prise de décision. D'une part, nous avons présenté l'intérêt de la procédure proposée pour la conception d'une base d'expériences formalisées, qui sera potentiellement une aide lors de la résolution de problèmes spécifiques. D'autre part, nous avons présenté les principales étapes développées pour extraire de nouvelles connaissances sous la forme de règles, à partir des bases de données disponibles dans les entreprises, qui nécessitent d'abord un processus de prétraitement et de préparation des données avant l'application des algorithmes d'extraction. Nous nous sommes intéressés ainsi à la création de nouvelles connaissances à l'aide d'une démarche de REx et à leur utilité possible, ce qui aborde le principal objectif de notre travail.

Tout au long de cette démarche générale proposée, nous avons donné une place privilégiée aux Graphes Conceptuels (GCs), formalisme de représentation des connaissances choisi pour structurer les différents types de connaissances dans la base de REx-ECD.

Finalement, nous allons approfondir dans le Chapitre III la dernière problématique abordée (II.3.3.2), i.e. l'analyse et l'exploitation des règles d'association dans la démarche REx-ECD, dans laquelle les graphes conceptuels, utilisés en particulier pour la représentation des règles extraites et la visualisation de leurs relations, facilitent la compréhension et l'interprétation des nouvelles connaissances par un utilisateur non-expert en *data mining*. Nous proposerons une approche guidée, pratique et utile afin d'évaluer et de comparer les règles extraites avant leur validation en suivant pas-à-pas une méthodologie structurée.

### III UNE DEMARCHE INTERACTIVE POUR LE POST-TRAITEMENT DANS LE PROCESSUS D'ECD

---

III.1	ROLE DE L'EXPERT DU DOMAINE DANS LE PROCESSUS D'ECD .....	64
III.2	INTERACTION ET COOPERATION ENTRE L'EXPERT DU DOMAINE ET L'EXPERT D'ECD.....	65
III.3	INTEGRATION DES ASPECTS D'INTERACTION HOMME-MACHINE DANS LE PROCESSUS .....	67
III.4	INTERPRETATION ET EVALUATION DES CONNAISSANCES EXTRAITES.....	68
III.4.1	EVALUATION OBJECTIVE : UNE ANALYSE D'UN POINT DE VUE STATISTIQUE.....	69
III.4.2	EVALUATION SEMANTIQUE : INTERPRETATION ET VISUALISATION DES RESULTATS.....	72
III.4.3	EVALUATION SUBJECTIVE : RECHERCHE DES TYPES SPECIFIQUES DE REGLES.....	82
III.5	BILAN .....	87

Au cours de ces dernières années, les efforts entrepris pour tenter de faire des progrès sur les processus d'extraction de règles d'association se sont concentrés sur l'amélioration de la performance algorithmique (notamment en termes de temps d'exécution et de mémoire utilisée) ; la phase de post-traitement du processus est souvent considérée comme avoir été particulièrement négligée (Wu et al., 2003).

Bien que l'extraction de règles d'association soit une technique de *data mining* largement utilisée, elle présente quelques difficultés : en particulier, un grand nombre de règles peuvent être créées, et toutes les règles découvertes peuvent ne pas être pertinentes pour supporter la prise de décision dans un domaine spécifique. Dans ce contexte, la phase de post-traitement devient nécessaire dans la procédure d'évaluation de l'utilité et de la fiabilité des résultats obtenus (Giudici, 2003), d'où l'importance d'une collaboration adéquate et d'un partage approprié de connaissances entre les différents experts potentiellement concernés.

Actuellement, l'interprétation et l'évaluation des résultats obtenus prennent donc de plus en plus d'importance dans les entreprises menant ce type d'action, afin de trouver des motifs ou des modèles potentiellement intéressants et utiles pour chaque problème spécifique. En effet, peu de systèmes de *data mining* développés dans les entreprises sont efficacement reliés avec les domaines d'application en tenant compte de leurs contraintes particulières (Pechenizkiy et al., 2005). Dans (Wu et al., 2003), les auteurs ont eux aussi souligné ce point : « *we see many papers proposing incremental refinements in association rules algorithms, but very few papers describing how the discovered association rules are used* ».

C'est pour cette raison que nous nous concentrons dans ce chapitre sur l'analyse et l'évaluation des résultats obtenus plutôt que sur l'optimisation de l'efficacité des algorithmes d'extraction existants.

Ce chapitre est dédié à la phase de post-traitement du processus d'ECD, liée à l'exploitation de la base de REx-ECD, et plus particulièrement de la base de règles extraites.

Par la suite, nous décrivons de manière progressive comment les règles découvertes peuvent être interprétées et évaluées par les utilisateurs. Plus particulièrement, nous proposons une approche conceptuelle structurée, pratique et interactive, afin d'interpréter, visualiser et filtrer les règles extraites, en suivant une méthodologie guidée qui vise à favoriser la collaboration entre experts de deux domaines différents : un expert d'extraction de connaissances (ou expert de *data mining*) et un expert métier, spécialiste du domaine d'application.

### III.1 RÔLE DE L'EXPERT DU DOMAINE DANS LE PROCESSUS D'ECD

En pratique, il est difficile de trouver un expert en extraction de connaissances (expert d'ECD) qui soit également un expert dans un domaine industriel considéré.

Les activités d'extraction de connaissances dans la démarche REx-ECD devraient donc être gérées par l'expert d'ECD mais centrées sur les besoins et les intérêts de l'utilisateur final, ici l'expert du domaine industriel. L'expert du domaine est pour nous la personne qui connaît le terrain et qui est responsable de la prise de décision dans l'entreprise. En revanche, l'expert d'ECD développe et améliore les techniques de *data mining* afin d'assurer une bonne alimentation en connaissances du système d'aide à la décision. De ce fait, une participation significative de l'expert du domaine industriel est suggérée dans le processus d'ECD.

Les connaissances de l'expert du domaine peuvent par exemple guider ou aider à la sélection des caractéristiques singulières et des modèles les plus significatifs, à la prise en compte des facteurs propres du domaine dans la modélisation, à la vérification des hypothèses du domaine, à la conception des mesures d'intérêt en incorporant des préoccupations propres de l'entreprise cible, et à l'évaluation des résultats issus du processus de *data mining* (Cao & Zhang, 2008).

Dans de nombreuses industries basées sur la connaissance, le manque de recherche sur les techniques d'ECD liées à des domaines d'application spécifiques a considérablement affecté l'utilité des activités de *data mining* (Pechenizkiy et al., 2008) (Wang & Wang, 2009), d'où notre intérêt pour la prise en compte des connaissances des experts du domaine, surtout lors de l'interprétation et de

l'évaluation des résultats obtenus, qui permettent la validation des éléments d'intérêt qui vont supporter la prise de décision dans l'entreprise.

Dans ce contexte d'évaluation, comment spécifier les caractéristiques intéressantes d'un modèle extrait ? Quatre notions peuvent aider à caractériser l'intérêt des modèles extraits (Marinica, 2010), notamment ici des règles d'association : la *validité*, la *nouveauté*, l'*utilité* et la *compréhension* par les utilisateurs. Les modèles extraits doivent en effet :

- valider l'ensemble des données analysées, et dans une certaine mesure de nouvelles données ;
- apporter de nouvelles connaissances à l'utilisateur ;
- être utiles afin de supporter la prise de décision ; et
- être facilement compréhensibles par cet utilisateur final.

Dans notre travail, nous nous concentrons plus particulièrement sur les deux dernières notions d'intérêt abordées, i.e. l'*utilité* et la *compréhension* par l'utilisateur expert du domaine des connaissances extraites. Ces notions sont associées au concept d'interactivité dans notre système REx-ECD, en insistant surtout sur le rôle indispensable des acteurs du domaine dans le processus (Ben Ayed et al., 2010) en vue d'adapter les techniques d'extraction de connaissances aux besoins industriels.

## III.2 INTERACTION ET COOPERATION ENTRE L'EXPERT DU DOMAINE ET L'EXPERT D'ECD

Nous décrivons dans cette partie des aspects importants de la coopération / collaboration entre experts du domaine industriel et experts d'ECD dans notre démarche, d'une part dans la définition des objectifs du processus d'extraction de connaissances, et d'autre part dans la phase de post-traitement des résultats issus de ce processus.

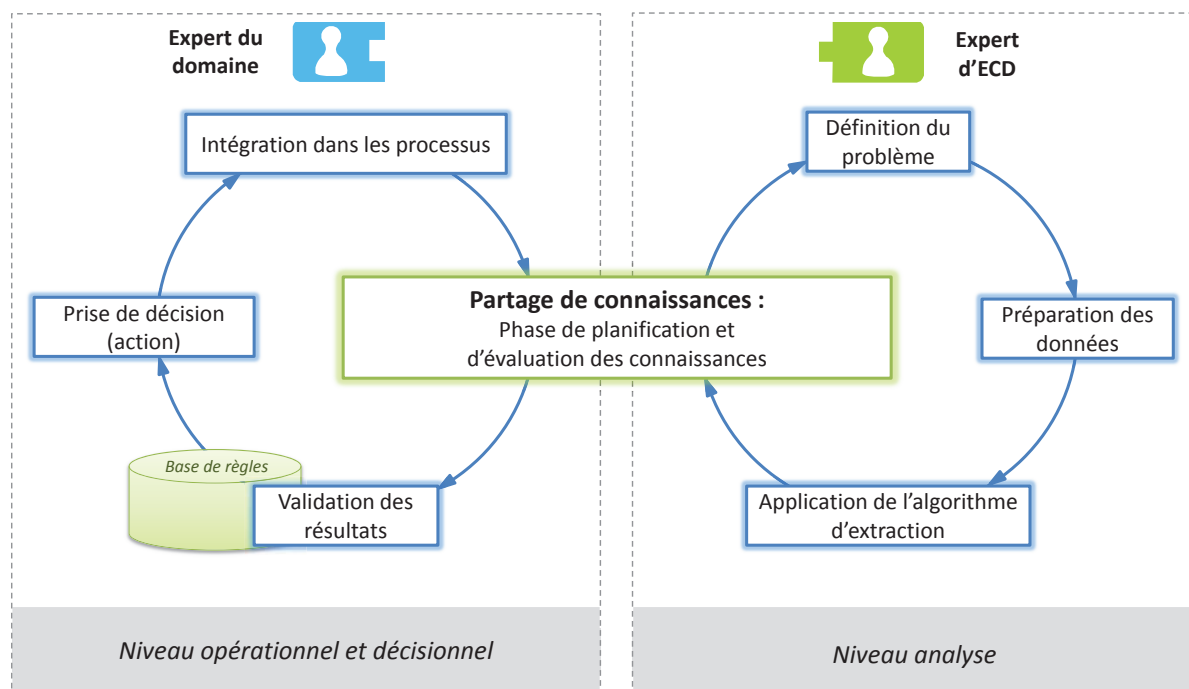
L'interaction entre l'expert du domaine et l'expert d'ECD est un processus de partage de connaissances particulier (Wang & Wang, 2008). C'est un aspect critique des applications de *data mining*, permettant de garantir la qualité des résultats obtenus.

Selon Hermiz (Hermiz, 1999), il existe quatre facteurs critiques dans le développement d'un projet d'extraction de connaissances :

- avoir clairement défini la problématique du domaine qui va être abordée et pour laquelle le processus d'ECD est un outil approprié ;
- s'assurer que les données disponibles pour l'analyse sont suffisantes et appropriées tant en terme de qualité que de quantité ;

- reconnaître que l'ECD est un processus qui contient de nombreux composants et dépendances ; il est donc difficile à gérer en totalité sans la collaboration d'un expert de *data mining* ;
- envisager de tirer profit du processus d'ECD quel que soit le résultat, et comprendre qu'il n'y a aucune garantie que tout projet de *data mining* soit un succès.

Une bonne implication de l'expert du domaine et la mise en œuvre d'un processus de partage des connaissances entre les experts peuvent améliorer l'efficacité du processus général d'ECD. En nous inspirant sur des travaux de Wang et Wang (Wang & Wang, 2008), nous suggérons un modèle conceptuel qui articule les connaissances entre les acteurs potentiellement impliqués (Figure III.1). L'intersection d'un cycle dédié à l'expert d'ECD et d'un autre dédié à l'expert du domaine est définie comme une phase de **partage de connaissances** conçue principalement pour analyser et évaluer les règles extraites ainsi que pour améliorer les futurs projets de *data mining*.



**Figure III.1. Modèle conceptuel proposé pour le partage de connaissances entre l'expert du domaine et l'expert d'ECD**

Tout d'abord, le cycle des opérations d'extraction des connaissances, lié notamment à l'expert d'ECD (chemin de droite sur la Figure III.1), concerne le niveau d'analyse du modèle. Dans ce contexte, une étape d'échange de connaissances entre les deux experts est d'abord réalisée par une **phase de planification**, afin de définir le problème initial ou l'objectif principal du cycle d'extraction de connaissances envisagé. Ensuite, les phases de prétraitement (i.e. la **préparation des données** incluant la sélection, le nettoyage et la transformation des données) et de *data mining* (i.e. l'**application de l'algorithme d'extraction** des règles) sont exécutées en fonction de cette planification initiale. Finalement, la phase de post-traitement (ou d'**évaluation des connaissances**) est envisagée afin d'interpréter et d'évaluer les résultats obtenus, en tenant compte de nouveau de



l'intervention de l'expert du domaine. Ce processus d'extraction de connaissances est ainsi guidé par l'expert d'ECD mais centré sur l'expert du domaine.

Le cycle lié à cet expert du domaine (chemin de gauche sur la Figure III.1) concerne le niveau opérationnel et décisionnel du modèle conceptuel proposé. Ici, il s'agit d'abord d'un traitement des résultats issus de la phase de *data mining*, d'une phase de **validation des résultats** tenant compte des connaissances de l'utilisateur du domaine, puis d'une **prise de décision** (stratégique ou opérationnelle) basée sur des connaissances filtrées. Finalement, une **intégration dans les processus** industriels des connaissances extraites peut être envisagée en vue d'une amélioration des processus existants. Pour achever le cycle, un retour d'expérience positif et/ou négatif doit être effectué vers l'expert d'ECD afin de formuler de nouveaux objectifs et d'évaluer la manière de présenter les résultats, pour améliorer les nouvelles tâches de *data mining* (i.e. un nouveau cycle d'extraction de connaissances).

Avant de présenter en détail la démarche proposée pour interpréter et évaluer les règles extraites lors de l'étape d'échange de connaissances entre les deux experts, nous pensons intéressant d'aborder dans la suite une thématique essentielle lors de la conception d'un système centré sur l'utilisateur final : l'interaction homme-machine (IHM), qui pourrait être mise en œuvre notamment lors de la phase d'évaluation des connaissances et de validation des résultats pour faciliter la réutilisation des règles.

### III.3 IMPORTANCE DES ASPECTS D'INTERACTION HOMME-MACHINE DANS LE PROCESSUS

Dans le développement d'un système d'aide à la décision interactif dédié à des utilisateurs experts du domaine l'interaction homme-machine (IHM) dans le processus est un aspect important pour que le système implémenté soit adapté au domaine d'application. Plus particulièrement, dans les processus d'extraction de connaissances, l'interaction entre l'utilisateur final et le système devient nécessaire pour garantir la qualité des résultats.

En effet, l'intégration de l'IHM dans un système d'aide à la décision simplifie, pour les utilisateurs finaux l'identification, l'exploration et la résolution des problèmes spécifiques, au travers du dialogue humain-ordinateur (Ben Ayed et al., 2010). En général, les modèles ou systèmes enrichis d'IHM sont donc centrés sur l'utilisateur, et sont caractérisés notamment par la priorité donnée aux aspects ergonomiques lors de la présentation des résultats. Toutefois, ces systèmes nécessitent une évaluation et une validation préalable à leur implémentation (Ben Ayed et al., 2010) (Millot & Debernard, 1993).

Dans la partie précédente, nous avons abordé l'interaction et la coopération entre humains (y compris les utilisateurs finaux et les experts de *data mining*) ; dans cette partie, nous présentons plutôt des aspects liés à l'interaction homme-machine lors de la présentation à l'utilisateur des résultats issus du processus d'ECD, i.e. l'interaction entre l'utilisateur final et le système d'extraction de connaissances (Figure III.2).

En effet, les connaissances extraites doivent être interprétées et évaluées à l'aide des experts afin de structurer les résultats et de valider les éléments d'intérêt proposés par l'algorithme de *data mining*. L'importance de prendre en compte explicitement et systématiquement l'utilisateur final conduit ainsi à une réflexion sur la manière de lui présenter les résultats finaux (via une interface conviviale bien adaptée), afin d'améliorer et d'optimiser le processus de filtrage lors de la réutilisation de ces résultats, et donc de prise de décision. La visualisation des connaissances extraites pourrait en effet aider l'utilisateur final à comprendre plus facilement l'information présentée et à tester ses différentes hypothèses afin de trier les résultats obtenus. Afin que cette étape du processus soit adaptée aux utilisateurs, nous suggérons d'utiliser une interface graphique touchant la visualisation de l'ensemble de règles extraites. Cette problématique se situe en effet à l'intersection de deux domaines de recherche : l'IHM et l'ECD (Chevrin et al., 2007). Elle est approfondie dans les sections III.4.2 et III.4.3.

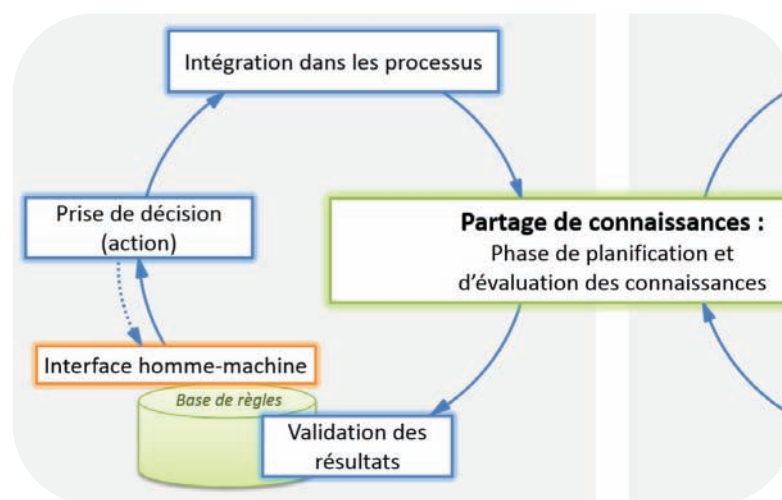


Figure III.2. L'intégration IHM dans notre approche conceptuelle

### III.4 INTERPRETATION ET EVALUATION DES CONNAISSANCES EXTRAITES

Le post-traitement dans un processus d'ECD est une phase nécessaire pour évaluer l'intérêt des règles extraites. Comme souligné dans la partie précédente, nous considérons qu'il ne devrait pas être automatique mais devrait être assuré par les experts de *data mining* et les experts du domaine, afin de garantir la pertinence du processus d'extraction dans une entreprise donnée.

La phase de post-traitement, concernant ainsi le partage de connaissances entre les experts et la validation des résultats, fait l'objet de cette partie.

Nous suggérons trois manières d'évaluer les règles d'association dans la démarche REx-ECD, en nous inspirant d'une classification présentée dans (Geng & Hamilton, 2006) :

- Une « **évaluation objective** » basée sur des mesures statistiques. Ici, nous allons d'abord interpréter le support et la confiance des règles extraites et les seuils *minsup* et *minconf* prédéfinis lors de l'application de l'algorithme *Apriori* (cf. Chapitre précédent).
- Une « **évaluation sémantique** » qui prend en considération les connaissances du domaine. Cette évaluation considère en effet la sémantique et les interprétations associées aux modèles et/ou motifs extraits, et permet d'évaluer leur utilité potentielle dans le domaine.
- Une « **évaluation subjective** » qui dépend surtout des objectifs, des croyances ou du point de vue de l'expert du domaine sur les données. Dans cette étape, l'expert du domaine pourrait exprimer des requêtes sur le type de connaissances produites, puis filtrer les résultats finaux du processus, susceptibles de l'aider dans ses prises de décision futures.

La démarche interactive proposée pour le post-traitement est ainsi contrôlée dans un premier temps par l'expert d'ECD, puis par l'expert du domaine d'application. Cette démarche prend en considération des travaux déjà décrits dans la littérature. Sa nouveauté résulte toutefois dans la manière d'interpréter, de visualiser et de filtrer les connaissances extraites, en tenant compte de plusieurs facteurs : l'analyse du support et de la confiance des règles, l'interprétation de la présence et absence de règles attendues dans les résultats, la visualisation des relations existantes entre les différentes règles extraites, la proposition d'une métarègle présentant les relations existantes entre les *items* détectés dans la base et regroupant l'ensemble des règles extraites dans une seule structure, l'interaction et la coopération entre l'expert du domaine et l'expert d'ECD dans le processus, et enfin l'interaction entre l'expert du domaine et le système d'aide à la décision (en utilisant un mécanisme de filtrage des résultats).

#### III.4.1 EVALUATION OBJECTIVE : UNE ANALYSE D'UN POINT DE VUE STATISTIQUE

Le mode traditionnel d'évaluation objective des connaissances (en tenant compte des mesures statistiques, ici le support et la confiance) est effectué généralement pendant l'extraction des règles d'association (Figure III.3) afin d'évaluer l'intérêt d'une règle d'un point de vue statistique. Nous nous concentrons dans cette partie sur l'analyse des seuils *minsup* et *minconf* prédéfinis et sur l'interprétation du support et de la confiance des règles obtenues.

Bien que d'autres mesures statistiques (également appelées mesures objectives ou indices de qualité) aient été proposées dans la littérature (voir section I.5.3.3), une évaluation objective des règles d'association est généralement effectuée en déterminant les règles qui ont un support et une confiance supérieurs ou égaux aux seuils prédéfinis (Agrawal et al., 1993). Par rapport aux autres mesures, un avantage important du support et de la confiance est leur grande intelligibilité, i.e. ces mesures sont facilement assimilables et interprétables par un utilisateur non expert (Lallich & Teytaud, 2004). Cependant, elles présentent des inconvénients et/ou des limites.

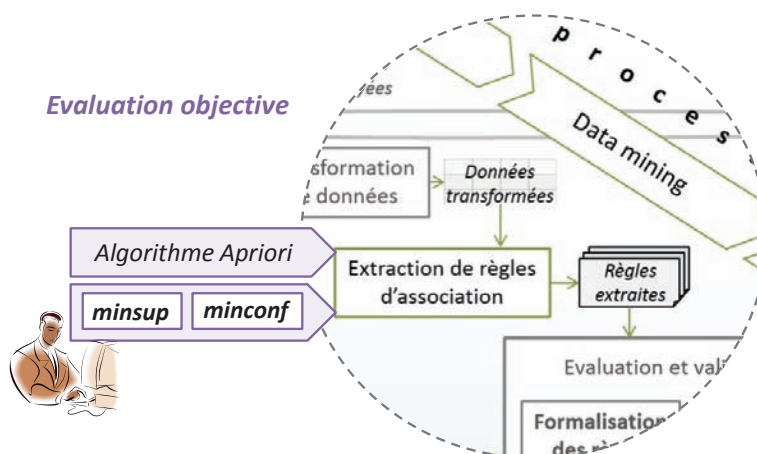


Figure III.3. Evaluation objective prise en compte lors de la phase de data mining dans notre démarche

Un inconvénient du support est que des connaissances intéressantes peuvent avoir un support inférieur au *minsup* alors que certaines d'entre elles peuvent avoir une très forte confiance. Toutefois, si l'on baisse le *minsup*, cela risque d'accroître considérablement le nombre d'*itemsets* fréquents et donc le nombre de règles d'association, qui peut être difficile à gérer lors du post-traitement. D'autre part, le principal inconvénient de la confiance est lié au fait que certaines règles extraites peuvent être non pertinentes même si elles ont une forte confiance, car ce paramètre ne prend pas en compte l'indépendance entre les attributs analysés. La Figure III.4, extraite de (Lallich & Teytaud, 2004), illustre ce point : la règle  $A \rightarrow B$  possède un support et une confiance élevés si l'on considère un ensemble  $T$  de 100 transactions, dont 80 transactions contiennent l'*item*  $A$ , 90 transactions contiennent l'*item*  $B$  et 72 transactions contiennent à la fois les *items*  $A$  et  $B$ . Ainsi, le support de la règle  $A \rightarrow B$  est de 72% et sa confiance de 90%. Nous remarquons dans ce cas que la confiance de la règle est égale à la probabilité de la partie conclusion, indépendamment de la partie hypothèse, c'est-à-dire,  $\text{confiance}(A \rightarrow B) = P(B|A) = P(B)$ . Ainsi, sur cet exemple, la règle  $A \rightarrow B$  n'est pas pertinente pour l'utilisateur car les attributs corrélés sont indépendants.

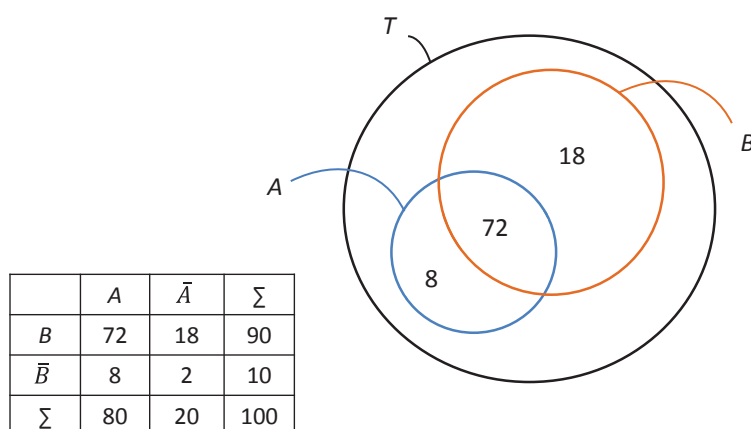


Figure III.4. Illustration des limites des mesures classiques Support-Confiance

Les seuils *minsup* et *minconf* doivent être définis avant l'application de l'algorithme *Apriori* (Agrawal & Srikant, 1994). En effet, ils correspondent à une première manière d'évaluer ou de filtrer les règles, sans garantir leur utilité et leur intérêt (Lallich & Teytaud, 2004). Or, choisir les niveaux optimaux de ces paramètres est une tâche ardue. Il a été souvent souligné que l'utilisateur a besoin d'une expertise spécifique du domaine afin de trouver le *minsup* et le *minconf* qui aboutiront à un nombre de règles optimal (Garcia et al., 2009). Comme alternative, un algorithme a été récemment proposé afin de découvrir un nombre donné de règles, ce qui devient utile par exemple lorsque l'utilisateur souhaite contrôler directement le nombre de règles extraites lors de l'application de l'algorithme : dans (Fournier-Viger et al., 2012), les auteurs ont proposé un algorithme pour extraire le « top-K » de règles d'association, où *k* est défini par l'utilisateur et concerne le nombre de règles d'association à trouver. Néanmoins, en raison de sa popularité, nous allons utiliser dans nos applications l'algorithme *Apriori* et tenter de gérer le problème d'une autre manière.

Analysons quelques cas particuliers. Définir par exemple un *minsup* très faible aboutira à une explosion combinatoire du nombre d'*itemsets* (ensemble d'*items*) candidats ; au contraire, un *minsup* très élevé empêcherait l'apparition des règles d'association contenant des *items* « rares » (non fréquents, inhabituels, exceptionnels ou sporadiques) mais potentiellement intéressants (Baesens et al., 2000). Autrement dit, si *minsup* = 0, chaque expérience analysée pourrait être exprimée par une règle différente : dans ce cas, dû au grand nombre de règles générées, aucune généralisation de la connaissance n'est effectuée. Des règles rares peuvent être parfois plus intéressantes que d'autres très fréquentes (et donc connues, même si quelques fois informulées) : cela dépend de la base de données considérée, du domaine d'application et/ou de l'objectif principal du processus d'extraction de connaissances. Il existe plusieurs algorithmes d'extraction de règles particulières qui se concentrent sur ce type de règles « rares ». Dans le domaine de la maintenance par exemple, si l'objectif est d'analyser les interventions urgentes, il faudra prendre en compte que les ordres de maintenance urgents sont généralement rares (même si cela dépend du type de maintenance effectuée). Ainsi, la définition d'un *minsup* élevé empêcherait de générer de la connaissance concernant ces interventions rares mais potentiellement intéressantes. Dans ce cas, il vaut mieux utiliser un algorithme d'extraction permettant d'obtenir ce type de règles.

Si le *minsup* est très élevé, seules les expériences qui combinent souvent le même ensemble d'*items* seront considérées. Si *minsup* = 1, une unique règle serait générée, sous la condition que toutes les expériences contiennent un même ensemble d'*items*. De ce fait, le *minsup* contrôle directement le nombre d'*itemsets* et de règles extraites, et par conséquent le degré de généralisation de la connaissance contenue dans ces règles.

Le *minconf* a un intérêt différent : il montre dans une certaine mesure la validité d'une règle, i.e. jusqu'à quel point la présence de la partie conclusion est liée à la présence de la partie hypothèse. Un *minconf* élevé permet de générer des règles très robustes, mais en pratique, ces règles sont généralement « évidentes » et bien connues des experts du domaine. Au contraire, les règles avec une faible confiance peuvent s'avérer insignifiantes, mais peuvent également exprimer des situations inhabituelles mais très intéressantes, suggérant des actions d'amélioration originales.

En pratique, un traitement efficace des attributs caractérisant les expériences demande de tester des seuils différents (voir Figure II.11. Illustration des résultats obtenus en fonction du *minsup* et du *minconf*), puisque, selon la base donnée analysée, certains seuils peuvent aboutir à un grand nombre de règles ou au contraire, à une quantité de règles insuffisante pour l'analyse.

Malheureusement, les travaux d'extraction de règles d'association se sont souvent limités à une évaluation classique ou objective (incluant le support et la confiance, ou d'autres mesures comparables), n'établissant l'intérêt d'une règle que d'un point de vue statistique, avec comme résultat un grand nombre de règles incohérentes, redondantes ou simplement des règles qui n'ont aucun intérêt du point de vue de l'utilisateur expert du domaine.

Étant donné que cette évaluation objective ne prend en considération qu'un aspect probabiliste, et n'assure pas l'élagage des règles non triviales, nous proposons une nouvelle approche afin de compléter cette manière d'évaluer les règles qui nous paraît insuffisante dans l'analyse des résultats obtenus (en termes de volume et de qualité).

Comme une alternative aux limites d'une « évaluation objective », nous suggérons ici de considérer également une « évaluation sémantique » et une « évaluation subjective » des règles extraites afin de faciliter leur compréhension par l'expert du domaine pour lui fournir en final des connaissances plus pertinentes.

### **III.4.2 EVALUATION SEMANTIQUE : INTERPRETATION ET VISUALISATION DES RESULTATS**

L'intégration de plus d'argumentation et d'analyse dans la phase de post-traitement devient nécessaire (Baesens et al., 2000) pour rendre les connaissances extraites plus utiles et compréhensibles par l'utilisateur final.

L'évaluation sémantique permet dans une certaine mesure d'évaluer l'intérêt d'une règle en tenant compte des connaissances du domaine. Toutefois, ces connaissances du domaine sont en général spécifiques, comportent de multiples facettes et peuvent être complexes ; il est donc difficile de donner ici des lignes directrices génériques. Un point particulier peut néanmoins être mis en avant : combiner l'« évaluation objective » (section III.4.1) et l'« évaluation sémantique » peut permettre de tirer des conclusions intéressantes et de diagnostiquer certains aspects de la cohérence et de la consistance de la base de données considérée.

Dans la littérature, il existe plusieurs techniques de post-traitement, parmi lesquelles celles qui se sont limitées à la visualisation des résultats, sans fournir vraiment d'aide à l'utilisateur pour évaluer l'utilité des règles. De ce fait, nous nous focalisons ici non seulement sur l'intérêt de la visualisation des résultats extraits, mais aussi sur leur interprétation en vue de tirer des conclusions ou même de suggérer des actions concrètes dans le domaine industriel.

Nous utilisons ici une évaluation sémantique des résultats, tout en tenant compte de l'ensemble des règles extraites, des connaissances du domaine (ici représentées sur le support ontologique des GCs), du modèle générique définissant une expérience et des connaissances de l'utilisateur (i.e. ses attentes, intérêts ou préférences). Ces quatre éléments constituent ainsi l'entrée de cette étape d'évaluation des règles extraites et de validation des résultats (Figure III.5).

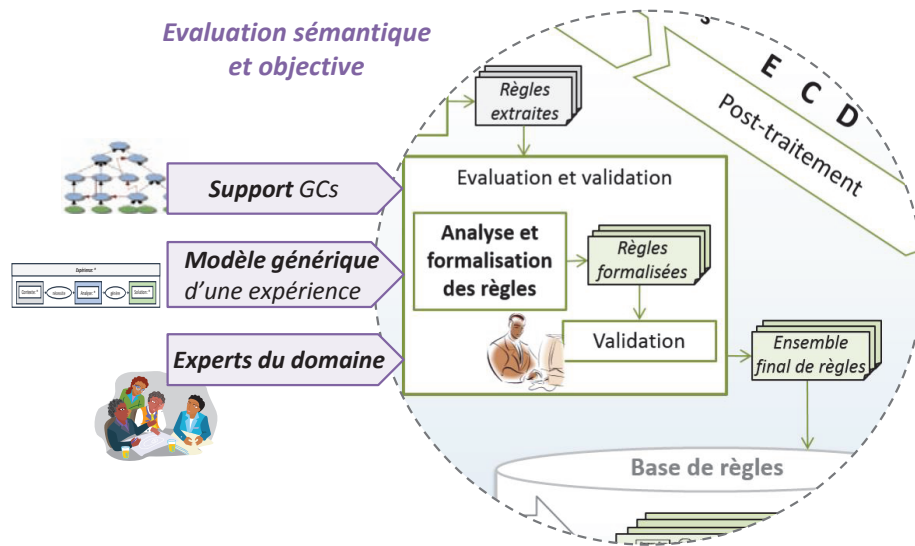


Figure III.5. Evaluation sémantique dans la phase de post-traitement

Pour cela, nous suggérons de parcourir étape-par-étape la démarche présentée ci-dessous pour mieux analyser, interpréter et visualiser les règles extraites :

- a) Analyser d'abord les « règles de base » (i.e. les règles n'impliquant que deux items).
- b) Associer une question à chaque attribut des items détectés contenus dans les règles.
- c) Exprimer de manière explicite le problème potentiel abordé par chaque règle en associant les questions correspondantes.
- d) Interpréter les mesures d'intérêt (ici de support et de confiance) de chaque règle.
- e) Analyser l'utilité possible des règles pour améliorer les processus industriels.
- f) Vérifier si la règle inverse est, ou devrait être présente dans les résultats obtenus.
- g) Considérer et analyser les règles plus complexes/spécifiques par rapport à celles de base au travers de trois opérations logiques, désignés ici comme **extension**, **permutation** et **jonction**<sup>5</sup>, puis parcourir de nouveau les étapes décrites ci-dessus.

<sup>5</sup> Ces opérations sont sémantiquement bien définies par des opérations des graphes conceptuels : généralisation/spécialisation, jointure et équivalence. L' « extension » constitue une forme de généralisation/spécialisation dans les GCs, la « jonction » une forme de jointure dans les GCs et la « permutation » est une forme légère d'équivalence dans les GCs tenant compte ici des mesures d'intérêt.



- h) Définir et représenter la structure globale de l'ensemble des règles extraites (en indiquant les relations entre règles de base et règles spécifiques).*
- i) Finalement, formaliser une « métarègle » afin de synthétiser l'ensemble de règles et décrire avec un niveau d'abstraction plus élevé toutes les relations entre les items détectés dans les règles extraites.*

Les étapes générales de la méthodologie sont détaillées dans la suite et illustrées à partir des résultats obtenus dans la Figure II.13. Une illustration plus détaillée, montrant mieux l'intérêt de cette méthodologie, sera présentée dans le chapitre suivant (Chapitre IV) lors de l'étude des cas réels abordés dans le domaine de la maintenance.

#### **III.4.2.1 Étapes de (a) à (f)**

Avant d'illustrer les premières étapes de la méthodologie concernant l'analyse et l'interprétation des résultats, nous présentons tout d'abord l'ensemble de règles extraites sur lequel nous allons illustrer toutes les étapes de la démarche.

Considérons ainsi le Tableau III.1 issu des résultats présentés sur la Figure II.13, conduisant à l'extraction de 10 *itemsets* et 12 règles d'association. Pour mieux illustrer les différentes étapes, nous avons remplacé les *items* détectés par l'algorithme *Apriori* par des valeurs d'attributs tirées du domaine de la maintenance de compresseurs (Weka, 2003). Les valeurs qui sont assignées à chaque *item* sont présentées ci-dessous :

A : « *Défaut*=filtre écrasé »

B : « *Action*=changer filtre »

C : « *Symptôme*=baisse de pression »

D : « *Cause*=poussières »

E : « *Cause*=humidité excessive »



Tableau III.1. Ensemble de règles extraites satisfaisant un *minsup* = 30% et un *minconf* = 50%

Règle #	Hypothèse	Sup (%)	Conf (%)
R1	<i>Cause</i> =humidité excessive → <i>Action</i> =changer filtre	30	100
R2	<i>Cause</i> =humidité excessive → <i>Défaut</i> =filtre écrasé	30	100
R3	<i>Action</i> =changer filtre → <i>Symptôme</i> =baisse de pression	40	50
R4	<i>Symptôme</i> =baisse de pression → <i>Action</i> =changer filtre	40	67
R5	<i>Défaut</i> =filtre écrasé → <i>Symptôme</i> =baisse de pression	40	57
R6	<i>Symptôme</i> =baisse de pression → <i>Défaut</i> =filtre écrasé	40	67
R7	<i>Défaut</i> =filtre écrasé → <i>Action</i> =changer filtre	50	71
R8	<i>Action</i> =changer filtre → <i>Défaut</i> =filtre écrasé	50	62
R9	<i>Cause</i> =humidité excessive, <i>Défaut</i> =filtre écrasé → <i>Action</i> =changer filtre	30	100
R10	<i>Cause</i> =humidité excessive, <i>Action</i> =changer filtre → <i>Défaut</i> =filtre écrasé	30	100
R11	<i>Défaut</i> =filtre écrasé, <i>Action</i> =changer filtre → <i>Cause</i> =humidité excessive	30	60
R12	<i>Cause</i> =humidité excessive → <i>Défaut</i> =filtre écrasé, <i>Action</i> =changer filtre	30	100

Nous considérons ici les huit premières règles établies par l'algorithme comme « règles de base », puisqu'elles ne mettent en relation que deux *items*. Pour chacune de ces règles, il peut être intéressant d'associer une question à chaque attribut des *items* analysés, afin de faciliter l'interprétation de l'utilité de la règle. Ci-dessous, nous proposons à titre d'exemple l'interprétation et l'analyse des règles R1 et R4 du Tableau III.1.

**R1** : *Cause*=humidité excessive → *Action*=changer filtre

*sup* = 30% *conf* = 100%

*Question associée* : Le problème abordé par cette règle peut être exprimé par une question associée abordant le lien possible entre les attributs mis en relation ; par exemple : Y a-t-il un lien entre l' « origine de l'intervention » (*cause*) et l' « action menée » ?

*Interprétation et utilité* : Ces deux *items* (« *cause*=humidité excessive » et « *action*=changer filtre ») sont souvent présents dans la base de données considérée (*sup* = 30%), i.e. 30% des interventions analysées contiennent ces deux *items*. L'*item* « *cause*=humidité excessive » s'accompagne toujours de l'apparition de l'*item* « *action*=changer filtre », i.e. les interventions ayant détecté que l'origine d'une défaillance sur un compresseur donné est une humidité excessive, entraînent toujours une même action pour remédier au dysfonctionnement constaté lors de l'intervention de maintenance : le changement du filtre (*conf* = 100%).

Par contre, cette action peut être liée à d'autres causes, étant donné que la règle inverse (*action*=changer filtre → *cause*=humidité excessive) est absente. En effet, elle n'a pas une confiance supérieure ou égale au *minconf* prédéfini (ici 50%). Cela nous permet de déduire que cette action de maintenance (changer filtre) peut traiter des problèmes dus à des causes différentes de celle de l'humidité excessive.

Cette analyse des règles (les règles présentes, mais aussi celles absentes des résultats obtenus), compte tenu des seuils *minsup* et *minconf* prédéfinis, peut permettre de mieux comprendre les

données analysées et d'identifier des incohérences ou inconsistances possibles dans les bases de données (par exemple dues à des fautes de frappe, erreurs de saisie ou encore anomalies dans la définition des attributs qui caractérisent les transactions). Considérons l'exemple précédent dans le domaine de la maintenance pour mieux illustrer l'intérêt de l'étape (f) dans la méthodologie proposée. Dans de nombreux cas réels, une cause donnée d'une défaillance devrait aboutir à une action spécifique de maintenance dans une situation donnée ; inversement, cette action de maintenance donnée ne devrait être effectuée que si cette cause est survenue. Dans ce cas, si une règle du type « *cause de la défaillance*=humidité excessive → *action de maintenance*=changer filtre » est générée, la règle inverse devrait également être générée si la base de données considérée a le même niveau de détail pour l'attribut « cause » et pour l'attribut « action » menée, puisqu'elle devrait avoir le même support et la même confiance dans le cas hypothétique où chaque cause aboutit à une action spécifique. Néanmoins, dans des cas pratiques, nous avons souvent constaté l'absence de règle inverse. Le fait qu'une action de maintenance puisse parfois traiter plusieurs problèmes distincts issus de causes différentes dans certaines bases de données peut être dû au fait qu'il n'y a pas un niveau de détail approprié dans la définition des causes et des actions associées. Cette anomalie peut être liée à la définition des attributs, par exemple à une différence dans la granularité choisie par le concepteur de la base de données (gestionnaire) et/ou les acteurs opérationnels au moment d'exprimer leurs expériences. La cause de la défaillance peut par exemple avoir des sous-causes pour l'acteur opérationnel, conduisant à différentes actions dans l'analyse.

**R4** : Symptôme=baisse de pression → Action=changer filtre

*sup* = 40% *conf* = 67%

*Question associée* : Y a-t-il un lien entre le « symptôme constaté » et l' « action menée » ?

*Interprétation et utilité* : Dans cette règle, les *items* « symptôme=baisse de pression » et « action=changer filtre » sont souvent associés dans la base de données (*sup* = 40%). 67% des interventions analysées ayant constaté comme symptôme une baisse de pression ont fait l'objet du changement du filtre, i.e. 33% des interventions contenant la valeur « baisse de pression » concernent une autre action menée. Dans ce cas, nous remarquons que la règle inverse est présente (règle R3 - Tableau III.1), et a le même support (car  $P(XY) = P(YX)$ ) ; toutefois en ce qui concerne la confiance de R3, il n'y a que 50% des interventions contenant l'*item* « action=changer filtre » qui contiennent l'*item* « symptôme=baisse de pression », les 50% restants concernant donc d'autres symptômes.

L'analyse de ce type de règle peut permettre aux acteurs du domaine de mieux préparer les interventions en fonction du symptôme constaté, en tenant compte par exemple des pièces de rechange les plus couramment utilisées pour traiter ce symptôme. En particulier, la confiance de la règle R4 (67%) indique qu'il peut être pertinent de prévoir un filtre pour l'intervention si une baisse de pression est constatée sur un compresseur.

### III.4.2.2 Etape (g)

Poursuivant l'exemple illustratif de la partie précédente, nous considérons maintenant les règles plus spécifiques par rapport à celles de base, i.e. les règles contenant plus de deux *items* (R9-R10-R11-R12). Ces règles peuvent être considérées comme des variantes des règles de base obtenues par les

trois opérations logiques évoquées : *extension* (de la partie hypothèse ou conclusion des règles), *permutation* (des *items* entre la partie hypothèse et conclusion des règles) et *jonction* (des parties hypothèse ou des parties conclusion des règles).

**Extension :** De nouveaux *items* peuvent être ajoutés aux parties hypothèse ou conclusion des règles de base, ce qui signifie que ces règles deviennent plus spécifiques et que les mesures d'intérêt peuvent être influencées par les nouveaux *items* ajoutés.

Nous allons présenter des exemples issus du Tableau III.1 pour illustrer cette opération. La règle R9 (*cause*=humidité excessive, *défaut*=filtre écrasé → *action*=changer filtre) est une extension de la partie hypothèse de la règle R1 (*cause*=humidité excessive → *action*=changer filtre). En effet, R9 n'ajoute que l'*item* « *défaut*=filtre écrasé » dans la partie hypothèse de R1, conservant toutefois le même support et la même confiance de la règle de base. Cette nouvelle règle, plus informative car elle concerne plus d'*items* tout en conservant le même support et la même confiance, devrait être retenue ici en lieu et place de la règle R1.

En outre, R9 est aussi une extension de R7 (*défaut*=filtre écrasé → *action*=changer filtre), obtenue en ajoutant l'*item* « *cause*=humidité excessive » dans la partie hypothèse. Dans ce cas, R9 a un support plus faible que R7 et une confiance plus forte, traduisant l'influence du nouvel *item* ajouté sur les mesures d'intérêt. Ici, nous considérons que les deux règles devraient être conservées, l'influence sur les indicateurs support et confiance de l'*item* « *cause*=humidité excessive » dans la nouvelle règle étant porteuse de connaissance.

**Permutation :** L'algorithme *Apriori* teste également des permutations des *items* entre la partie hypothèse et la partie conclusion des règles extraites. Dans ce cas, le support des règles est le même (car le support est une mesure symétrique) ; seule, la confiance peut varier du fait que cette mesure dépend de la probabilité de l'*item* (ou *itemset*) situé dans la partie hypothèse.

Pour illustrer la permutation entre règles, prenons comme exemple la règle R4 (analysée dans la partie précédente), qui concerne une permutation de R3 (*action*=changer filtre → *symptôme*=baisse de pression) (Tableau III.1). L'*item* « *symptôme*=baisse de pression » qui est dans la conclusion de R3 et l'*item* « *action*=changer filtre » qui est dans l'hypothèse ont été permutés dans la règle R4. Concernant la confiance, chaque règle fournit une information différente à l'utilisateur. Par exemple, la règle R3 indique que 50% des interventions ayant changé le filtre du compresseur ont été dues à une basse de pression. En outre, nous pouvons déduire de R4 que 33% des interventions ayant comme symptôme une base de pression concernent une autre valeur de l'attribut « *action* » (la confiance de la règle étant de 67%). D'autres permutations identifiées dans cet exemple sont : R6 (permutation de R5) et R8 (permutation de R7).

Nous ne pouvons pas préciser quelle règle devrait être retenue par rapport aux autres (chaque règle a un intérêt particulier) ; nous donnons néanmoins ici quelques pistes pour mieux interpréter la confiance des règles extraites, qui permet d'analyser l'influence potentielle des *items* de la partie hypothèse sur ceux de la partie conclusion.

**Jonction :** Nous pouvons remarquer que certaines règles combinent les parties conclusion de règles qui ont la même partie hypothèse. Inversement, des règles combinent les parties hypothèse de règles qui ont la même partie conclusion.

La règle R9 (*cause*=humidité excessive, *défaut*=filtre écrasé → *action*=changer filtre) du Tableau III.1 combine par exemple les parties hypothèses des règles R1 (*cause*=humidité excessive → *action*=changer filtre) et R7 (*défaut*=filtre écrasé → *action*=changer filtre), menant à la même partie conclusion. D'autres jonctions identifiées sont : la règle R10 combine la partie hypothèse des règles R2 et R8 ; R12 combine pour sa part les parties conclusions des règles R1 et R2, qui ont la même partie hypothèse.

Dans la partie concernant l'opération d'« extension », nous avons tenté d'interpréter le fait d'ajouter un nouvel *item*, soit dans la partie hypothèse, soit dans la partie conclusion d'une nouvelle règle. Dans cette partie (opération de « jonction »), nous avons présenté à titre illustratif les jonctions identifiées, qui fournissent à l'utilisateur des informations complémentaires sur l'origine des règles complexes/spécifiques générées (ici par exemple les règles de base qui les constituent) et peuvent le conduire à éliminer des règles de base de sa base de connaissance.

### III.4.2.3 Etape (h)

Définir et représenter la structure globale des règles peut faciliter leur compréhension, par l'exploration visuelle de l'ensemble de règles, et procurer ainsi une aide à l'utilisateur final lors de la validation et de l'exploitation des résultats (interaction homme-machine).

Différentes techniques de visualisation ont été développées pour faciliter la compréhension des règles découvertes, souvent fournies sous forme de simples listes textuelles à la sortie des algorithmes d'extraction. Les méthodologies utilisées pour visualiser ces règles d'association sont généralement basées sur des tableaux et des graphiques.

Nous avons présenté tout d'abord dans notre exemple une première visualisation de l'ensemble de règles extraites sur la forme de tableau (Tableau III.1), toutefois un inconvénient de cette représentation est la difficulté d'interpréter l'ensemble des règles lorsqu'elles sont en nombre important. Comme alternative, plusieurs méthodes graphiques ont été proposées dans la littérature (Blanchard, 2005), afin de présenter la base de connaissances d'une manière plus adaptée et ergonomique, facilitant ainsi la compréhension et la perception de l'information par l'expert du domaine d'application.

Nous présentons ci-dessous des méthodes de la littérature visant à représenter l'ensemble de règles extraites sous forme visuelle :

- La *représentation matricielle* permet la visualisation des règles d'association et des mesures de qualité (Couturier, 2005) par des objets 2D sous forme de matrices, ou 3D sous forme d'histogrammes.
- La *représentation par graphes* présente l'ensemble de règles sous la forme d'un graphe orienté, dans lequel les nœuds et les arcs sont utilisés pour représenter (respectivement) les *itemsets* détectés et les règles extraites (Kuntz et al., 2000). Cette méthode graphique est considérée comme étant plus intuitive que la matricielle, toutefois il faut tenir compte du fait que la propriété de transitivité des graphes ne peut pas s'appliquer au contexte des règles d'association, leur présence étant conditionnée par les mesures d'intérêt.

- La *visualisation des règles en 3D* (en utilisant diverses métaphores) et en *réalité virtuelle* (Blanchard, 2005) représente des objets répartis dans l'espace qui permettent une exploration très intuitive des données par l'utilisateur. Ces méthodes sont cependant hors du contexte de la démarche REx-ECD proposée dans ce mémoire.

Dans notre approche conceptuelle, pour des raisons de cohérences des outils manipulés, nous proposons de nouveau une représentation par graphes, qui fournira une représentation améliorée des motifs extraits et facilitera l'exploration visuelle et la compréhension des résultats par l'expert du domaine. Plus particulièrement, nous nous concentrons sur le formalisme des GCs (formalisme de représentation proposé pour structurer la base de REx-ECD), dotés d'une logique et d'une sémantique formelle. Les GCs vont ainsi supporter la visualisation des résultats, en synthétisant l'ensemble complet des règles extraites et des relations entre elles, mais aussi en suggérant la formalisation de chaque règle dans une métarègle.

Dans cette étape de la méthodologie, nous fournissons une vue globale de l'ensemble de règles d'un niveau général (règles qui ont le moins d'*items*) à un niveau plus spécifique (règles comprenant de nombreux *items*) à l'aide d'un modèle graphique (GCs), qui procure une certaine facilité de lecture des résultats obtenus. Les règles sont ainsi regroupées en deux niveaux d'abstraction : les règles au niveau supérieur fournissent un aperçu général de la connaissance extraite (i.e. les règles de base) tandis que les règles à un niveau inférieur correspondent à des règles plus spécifiques (en effet, plus il y a d'*items*, plus la règle est spécialisée, et plus bas elle sera représentée dans la structure). Ceci fournit à l'utilisateur la possibilité de visualiser les liens existants entre les règles extraites et d'aborder la base de connaissances à un niveau général, puis de la détailler selon ses intérêts du moment.

Nous illustrons ici une structure globale de l'ensemble des règles extraites présentées dans le Tableau III.1, en tenant compte des différents liens identifiés entre les règles. Les huit règles de base sont donc au niveau supérieur, et les règles plus spécifiques sont au niveau inférieur (Figure III.6).

Chacune des règles extraites est représentée par des types de concept contenus dans *TC* (rectangles), les mesures de support et confiance étant représentées comme une instance individuelle dans chaque règle, et les liens entre les règles étant représentés par des relations contenues dans *TR* (ovales), décrivant les opérations logiques définies dans la partie précédente.

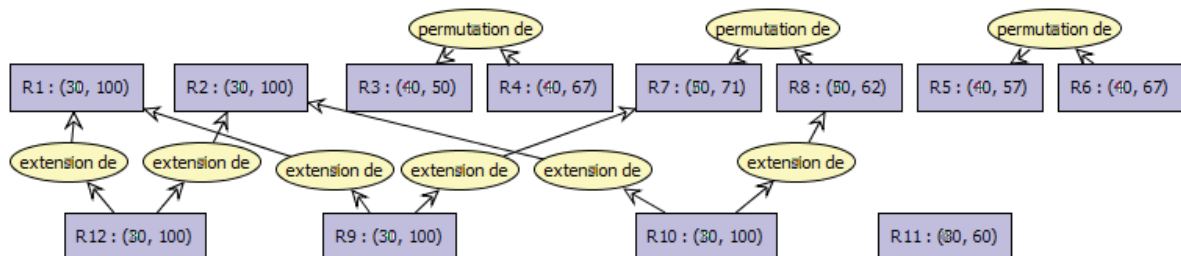


Figure III.6. Liens entre les différentes règles extraites (*sup, conf*)

Nous pouvons remarquer que la relation de « jonction » n'est pas explicitée dans la Figure III.6, puisque dans notre exemple une règle qui est une extension de deux règles de base différentes traduit une jonction de ces deux règles. Par exemple, la règle R12 est une extension de R1 (obtenue en ajoutant l'*item* « *défaut*=filtre écrasé » dans la partie conclusion) et de R2 (obtenue en ajoutant « *action*=changer filtre » dans la partie conclusion) : R12 (*cause*=humidité excessive → *défaut*=filtre écrasé, *action*=changer filtre) est donc une jonction des règles R1 (*cause*=humidité excessive → *action*=changer filtre) et R2 (*cause*=humidité excessive → *défaut*=filtre écrasé). Nous pouvons voir également sur la Figure III.6 qu'une règle spécifique (règle R11) n'est pas issue d'une règle de base, du fait du *minconf* prédéfini ici.

La formation de cette structure à l'aide des GCs complète (mais ne remplace pas) l'information fournie dans le tableau initial, donne une première organisation des connaissances dans la base de règles et contribue à l'amélioration de l'intelligibilité de l'ensemble de règles. L'utilisateur peut ainsi prendre conscience des complémentarités des règles et cibler l'ensemble final de règles qui lui paraît le plus utile par rapport à ses besoins. Par exemple, il serait intéressant de porter une attention particulière aux règles les plus spécifiques dans un contexte de résolution d'un problème particulier ; d'autre part, si l'objectif principal est d'améliorer le système, il est sans doute pertinent de s'intéresser d'abord aux règles de base.

#### III.4.2.4 Etape (i)

Finalement, nous proposons de formaliser une *métarègle* décrivant les relations existantes entre tous les *items* contenus dans les règles extraites. Les relations (contenues dans *TR*) traduisent une liaison directe entre les *items* (ici des concepts contenus dans *TC*) des parties hypothèse et conclusion des règles. Tout d'abord, pour formaliser chacune des règles extraites, nous avons construit des graphes conceptuels simples qui relient les *items* de l'hypothèse et de la conclusion par la relation « lié à », et les *items* contenus dans une même partie (hypothèse ou conclusion) par la relation « et ». Ensuite, pour formaliser la métarègle proposée, nous avons réuni dans un nouveau graphe les différents graphes conceptuels représentant chacune des règles grâce à l'« opération de jointure »<sup>6</sup> définie dans les graphes conceptuels. Nous avons représenté sur les arcs d'entrée (ou arcs amont) de chaque relation « lié à » la confiance de chaque règle. Toutes les règles extraites seront ainsi contenues dans la métarègle, qui procure une vue très synthétique sur la base de connaissances.

Nous présentons dans la Figure III.7 la métarègle issue des résultats du Tableau III.1, agrégeant les 12 règles extraites.

---

<sup>6</sup> *Opération de jointure* : Si un concept  $[c]$  dans un graphe conceptuel  $U$  est identique (type et marqueur) à un concept  $[d]$  dans un graphe  $V$  alors  $U$  et  $V$  peuvent être joints pour former un nouveau graphe  $W$ , en supprimant  $[d]$  et en liant à  $[c]$  toutes les relations qui étaient liées à  $[d]$ .



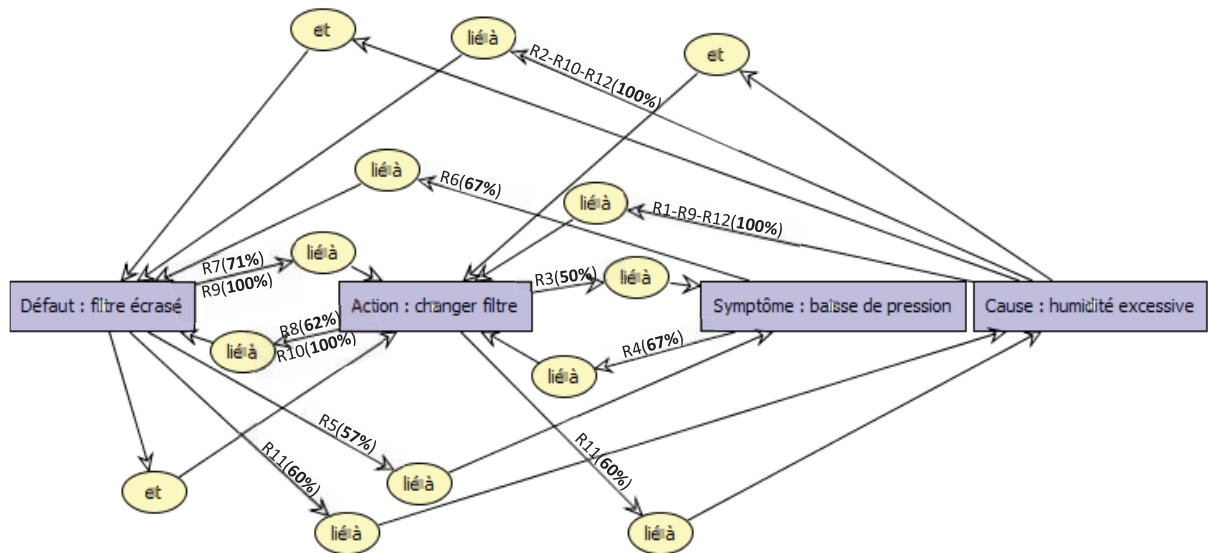


Figure III.7. Métarègle regroupant les items des règles extraites et leurs liens

Grâce à l'opération de projection des GCs (section II.2.4), nous pouvons vérifier la connaissance représentée dans la nouvelle structure. Nous illustrons par exemple le cas de la règle R12 contenue dans la métarègle, en projetant R12 sur la métarègle (Figure III.8). Il existe en effet une projection du GC de la règle R12 vers le GC de la métarègle (toutes les règles étant contenues dans ce graphe). Les concepts et les relations en couleur noire sur la métarègle de la Figure III.8, correspondent aux projections trouvées.

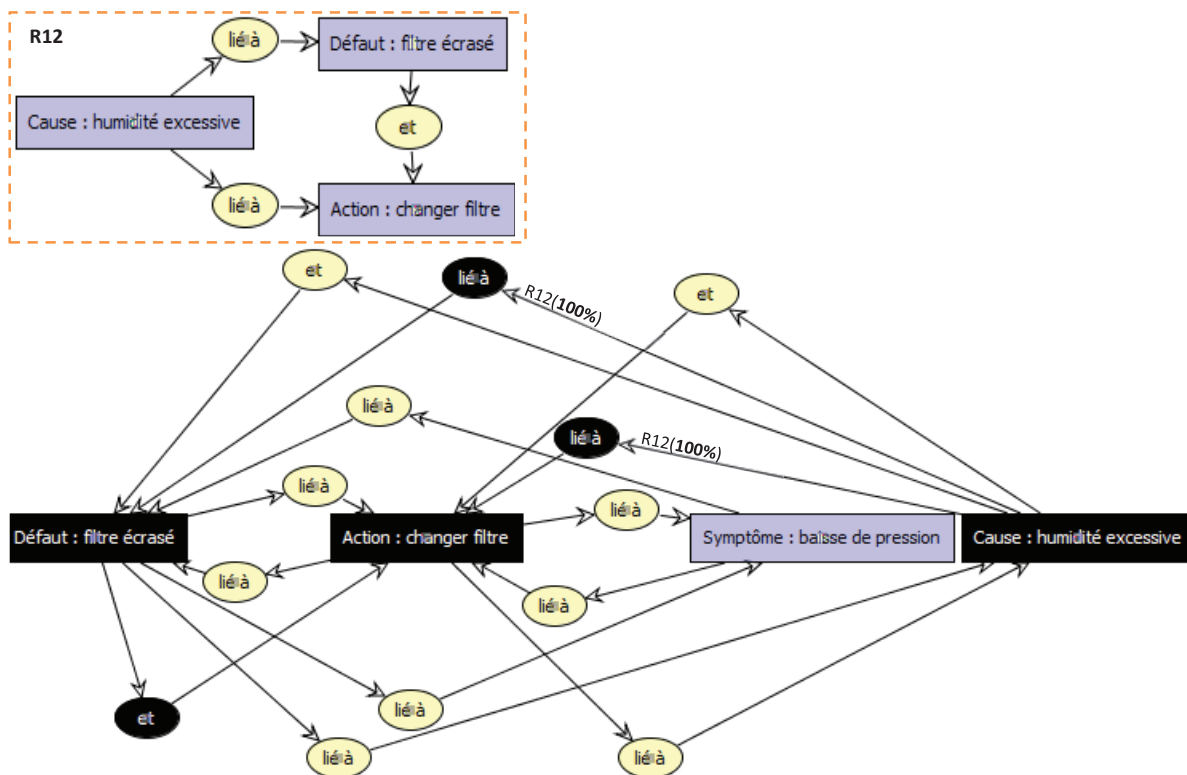


Figure III.8. Illustration de l'opération de projection de la règle R12 dans la métarègle

Après avoir considéré les GCs pour la représentation des liens entre les règles (*étape h*) et des liens entre les *items* des règles (*étape i*), nous allons de nouveau les utiliser dans la partie suivante comme mécanisme de raisonnement pour supporter le filtrage de l'ensemble des règles extraites, en tenant compte des attentes de l'expert du domaine.

### III.4.3 EVALUATION SUBJECTIVE : RECHERCHE DES TYPES SPECIFIQUES DE REGLES

Dans notre démarche, la participation de l'utilisateur expert du domaine doit permettre l'intégration de facteurs subjectifs lors de la validation et l'exploitation des résultats. Dans cette partie, l'objectif principal est d'aider l'expert du domaine à distinguer et filtrer les connaissances potentiellement utiles, intéressantes et/ou pertinentes parmi l'ensemble des règles d'association extraites.

L'évaluation subjective est avant tout liée à la recherche de types spécifiques de règles selon les attentes de l'expert du domaine. Dans un premier temps, la structuration des règles facilite l'exploration visuelle, puis elle assiste l'expert dans l'étape de filtrage de l'ensemble final des règles, phase pendant laquelle l'interaction entre l'utilisateur et le système REx-ECD est nécessaire. Il est indispensable de comprendre ce que l'utilisateur recherche dans les données pour pouvoir représenter ses attentes et filtrer les règles correspondant à ses attentes.

Dans notre processus d'ECD, la connaissance de l'objectif des décisions à supporter n'a pas été considérée comme un prérequis à l'application de l'algorithme d'extraction, à la différence d'autres techniques d'extraction de règles qui contraignent le nombre d'*items* et/ou déterminent quels *items* vont dans la partie hypothèse ou conclusion.

Les algorithmes d'extraction de connaissances comme *Apriori* (Agrawal & Srikant, 1994) permettent en effet d'extraire différents types de règles contenus dans une base de données, notamment des règles qui peuvent être attendues (en sachant qu'un utilisateur expert du domaine dans une situation donnée a souvent une idée sur le type de règle qu'il attend), ainsi que d'autres qui peuvent être complètement inattendues par l'utilisateur.

Ces règles inattendues peuvent être aussi très intéressantes, et fournir à l'utilisateur de nouvelles connaissances (Silberschatz, 1995).

#### III.4.3.1 Mise en contexte

Il existe dans la littérature différentes techniques pour effectuer une évaluation subjective permettant le filtrage des règles extraites.

Par exemple, la technique des « *templates* » (ou modèles) utilise des contraintes syntaxiques pour filtrer les règles qui pourraient intéresser ou pas l'utilisateur (Klemettinen et al., 1994). En effet, les *templates* spécifient quels attributs doivent être présents dans la partie hypothèse et dans la partie conclusion. Il existe donc des *templates* inclusives pour trouver des règles intéressantes et des *templates* restrictives pour trouver des règles inintéressantes. Cependant, cette proposition est considérée comme manquant de souplesse car tous les éléments d'une règle extraite doivent être une instance de tous les éléments du *template* pour que la règle soit retenue.



D'autre part, Silberschatz et Tuzhilin (Silberschatz & Tuzhilin, 1996b) examinent l'intérêt d'une règle selon deux critères subjectifs : l'actionnabilité (*actionability*) et le caractère inattendu (*unexpectedness*). La notion d'actionnabilité évalue si une règle peut être appliquée dans un domaine donné, le caractère inattendu exprimant la nouveauté de la règle pour l'utilisateur. Pour évaluer ce caractère inattendu, les auteurs classifient les croyances de l'utilisateur en deux groupes : fortes croyances et faibles croyances, associant ainsi un niveau aux connaissances de l'utilisateur sur le domaine.

Sahar (Sahar, 1999) fournit à l'utilisateur un ensemble de règles d'association obtenues et lui demande de classer ces règles comme étant vraies/fausses et intéressantes/inintéressantes. On peut ainsi les filtrer pour ne retenir que les règles intéressantes.

Un autre exemple est la technique des « *general impressions* » (impressions générales) (Liu et al., 1997). Les auteurs font ici une classification de la connaissance de l'utilisateur. D'une part, les impressions générales concernent des connaissances vagues que peut avoir l'utilisateur selon ses connaissances du domaine ; d'autre part, l'utilisateur peut avoir des connaissances précises concernant les règles extraites. Ces connaissances sont comparées syntaxiquement avec les règles extraites (parties hypothèse et conclusion), et les règles intéressantes sont retenues.

Une étude de nombreuses techniques subjectives basées sur les connaissances/attentes de l'utilisateur a été proposée dans (Marinica, 2010), montrant qu'il existe de nombreuses mesures subjectives et formalismes utilisables pour représenter et filtrer les connaissances extraites (*templates*, croyances, requêtes, méta-règles, taxonomies et ontologies en sont des exemples).

Nous avons, dans les sections II.3.2.1 et III.4.2, considéré les mécanismes de visualisation du formalisme des GCs pour la représentation des différents types de connaissances considérées dans notre approche (support ontologique, expériences et règles). Pour la détermination des modèles/règles les plus intéressants parmi ceux qui sont triviaux ou déjà bien connus par les experts du domaine, nous allons considérer des requêtes de l'utilisateur sur l'ensemble de règles extraites à l'aide du mécanisme de raisonnement de ce formalisme de représentation de connaissances : l'opération de projection.

#### III.4.3.2 Attentes de l'utilisateur

La représentation visuelle de l'ensemble de règles d'association à l'aide des GCs facilite l'interaction avec l'utilisateur et rend particulièrement facile le processus de modélisation d'une requête, puis le processus de filtrage selon la forme de la règle et/ou la présence impérative de certains *items*/concepts ou de leurs spécialisations dans *TC* (hiérarchie des types de concepts).

Une attente de l'utilisateur (qu'on appellera ici requête *Q*) se traduit par un squelette de règle, décrivant la structure des règles a priori intéressantes pour l'utilisateur. Un exemple de requête, dans laquelle l'utilisateur recherche parmi les règles extraites celles de la forme  $A \rightarrow B$ , est décrit sur la Figure III.9 à l'aide d'un graphe conceptuel. Dans ce modèle de règle ou type de règle attendue, l'utilisateur décrit les attributs (ou valeurs d'attributs) qui devraient être inclus dans l'hypothèse et ceux qui devraient apparaître dans la conclusion de la règle. Ainsi, parmi l'ensemble de règles extraites de notre exemple illustratif, l'utilisateur cible ici celles de la forme [Cause : \*]  $\rightarrow$  (lié

à)→[Action : \*], i.e. des règles contenant l'attribut « cause » dans la partie hypothèse et l'attribut « action » dans la partie conclusion, sans restreindre ces attributs à des valeurs spécifiques.

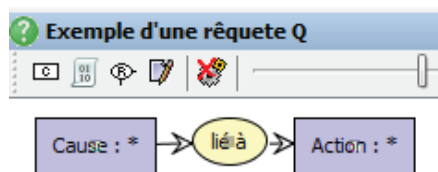


Figure III.9. Illustration d'une requête (Q) de l'utilisateur exprimée par un GC

### III.4.3.3 Mécanisme requête/réponse dans les graphes conceptuels

Dans le mécanisme requête/réponse envisagé, la recherche des règles qui répondent aux attentes de l'utilisateur est facilitée par l'opération de projection définie sur les GCs (section II.2.4), qui traduit une relation de subsomption entre graphes. Le fait que le même langage (i.e. les GCs) puisse être utilisé pour définir l'interface (pour construire la requête) et au niveau opérationnel (pour le traitement et la recherche des projections) rend transparente la structure logique de l'information, ce qui facilite la compréhension et l'interprétation des résultats par l'utilisateur (Mugnier, 2000).

Plus particulièrement, les relations de spécialisation/généralisation sont des notions fondamentales pour pouvoir raisonner avec les graphes. La relation de subsomption définit les attributs ou *items* qui appartiennent à une classe de concepts. Autrement dit, il est possible de déterminer si la description d'un concept donné est plus générale que la description d'un autre concept selon la connaissance ontologique définie dans le support (*S*). L'implication sémantique de la relation de subsomption est l'héritage des propriétés du concept parent (qui subsume un autre concept) vers le concept enfant (qui est subsumé) dans la hiérarchie des types de concepts *TC*. En effet, l'opération de projection concerne une notion algorithmique clé pour le raisonnement sur les GCs, puisqu'elle correspond à la procédure de subsomption ou à l'implication logique si l'on considère les formules logiques associées aux GCs.

Dans la partie précédente, nous avons représenté les règles extraites à l'aide des GCs en tenant compte des *items*/concepts définis dans le support ontologique, qui ont aussi permis de représenter les expériences. Nous proposons maintenant de représenter aussi les attentes de l'utilisateur à l'aide de ce formalisme.

Pour effectuer l'opération de projection, un mécanisme de requête/réponse va donc rechercher les projections entre le graphe « requête » (attente de l'utilisateur) et les graphes « réponse » (règles extraites formalisées) afin de trier des ensembles de règles potentiellement intéressantes pour l'utilisateur. Cette manière d'évaluer les règles est considérée comme subjective, puisqu'elle dépend du type de règle que l'utilisateur souhaite trouver dans la base de règles.

D'un point de vue algorithmique, le problème de projection qui consiste à déterminer si un graphe donné peut être projeté sur un autre graphe a en général un temps de calcul polynomial non-déterministe (Mugnier, 1995). Certains cas polynomiaux sont obtenus en restreignant la forme des graphes dans les applications pratiques, en particulier avec un temps polynomial qui dépend du choix

du graphe requête (Baget & Mugnier, 2002). Ainsi, la projection d'un GC acyclique dans un autre graphe est polynomiale (Chein & Mugnier, 2008).

#### III.4.3.4 Les différents types de règles découvertes

Il existe différents types de règles potentiellement intéressantes pour l'utilisateur, suivant ses attentes. Soit  $R$  un ensemble de règles d'association découvertes et  $Q$  une requête de l'utilisateur. En tenant compte de la structure des règles, une requête  $Q$  d'un utilisateur va trier l'ensemble de règles  $R$ . A ce propos, dans (Liu et al., 1999) les auteurs suggèrent de distinguer quatre classes de règles potentiellement intéressantes, que nous allons considérer dans notre travail lors de l'opération de projection.

##### Règles conformes

Une règle extraite  $R_i \in R$  est conforme à la requête  $Q$  de l'utilisateur si les deux parties hypothèse et conclusion de  $R_i$  peuvent être dérivées de  $Q$ , i.e. si  $R_i$  correspond à une projection de  $Q$ . Les règles conformes sont des règles dérivées ayant à la fois les parties hypothèse et conclusion en cohérence avec l'attente de l'utilisateur.

Poursuivant l'exemple des sections précédentes, nous illustrons sur la Figure III.10 une règle conforme obtenue par la projection de la requête de l'utilisateur (Figure III.9) vers l'ensemble de règles extraites. Nous remarquons que la connaissance représentée par la règle de base R1 peut être déduite du graphe requête (attente de l'expert du domaine) en tenant compte des hiérarchies  $TC$  et  $TR$ , étant donné que les attributs « cause » et « action » sont contenus dans la règle (mais ici avec des instances spécifiques).

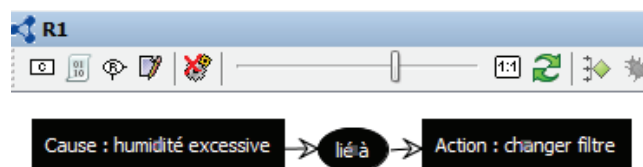


Figure III.10. Illustration d'une règle conforme à la requête de l'utilisateur

##### Règles avec conclusion inattendue

Une règle  $R_i \in R$  a une conclusion inattendue par rapport à  $Q$  si la partie hypothèse de  $R_i$  correspond à une projection de l'hypothèse de  $Q$ , mais pas la partie conclusion. On considère également dans cette catégorie des règles qui sont conformes à la requête de l'utilisateur  $Q$ , mais dont la partie conclusion fournit plus d'information par rapport à la requête. Les règles dont la conclusion est inattendue peuvent inclure des règles incohérentes avec la connaissance existante.

Parmi l'ensemble de règles extraites de l'exemple illustratif, les règles R2 et R12 concernent cette classe de règle (Figure III.11). D'une part, nous pouvons remarquer que l'hypothèse de la règle R2 (partie gauche de la Figure III.11) est conforme à l'hypothèse de la requête (Figure III.9), mais pas la

partie conclusion. D'autre part, la règle R12 (partie droite de la Figure III.11) admet une projection de  $Q$ , mais sa partie conclusion fournit une information complémentaire.

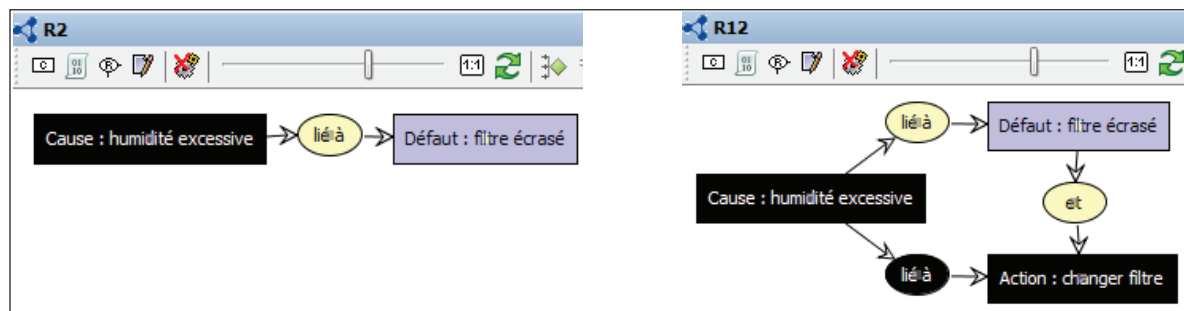


Figure III.11. . Illustration de règles avec partie conclusion inattendue

### Règles avec hypothèse inattendue

Une règle extraite  $R_i \in R$  a une hypothèse inattendue par rapport à  $Q$  si la partie conclusion de  $R_i$  correspond à une projection de la partie conclusion de  $Q$ , mais pas la partie hypothèse. On considère aussi dans cette catégorie des types de règles qui sont conformes à la requête  $Q$ , mais dont la partie hypothèse fournit plus d'information, menant à la même partie conclusion. Les règles dont l'hypothèse est inattendue peuvent montrer d'autres hypothèses qui peuvent conduire à un même résultat.

Après une projection de la requête de l'utilisateur (Figure III.9) vers l'ensemble de règles extraites, nous illustrons sur la Figure III.12 deux exemples de règles retenues dont l'hypothèse est inattendue. La règle R4 (partie gauche de la Figure III.12) a une partie hypothèse inattendue par rapport à la requête de l'utilisateur, tandis que la conclusion de cette règle correspond à une projection de la conclusion de  $Q$ . Nous pouvons remarquer aussi que la règle R12 (partie droite de la Figure III.12), est conforme à la requête, mais que sa partie hypothèse fournit plus d'information (dans ce cas, le défaut constaté). La règle R9 doit donc être considérée dans cette catégorie. Une autre règle extraite retenue dans cette catégorie est la règle R7.

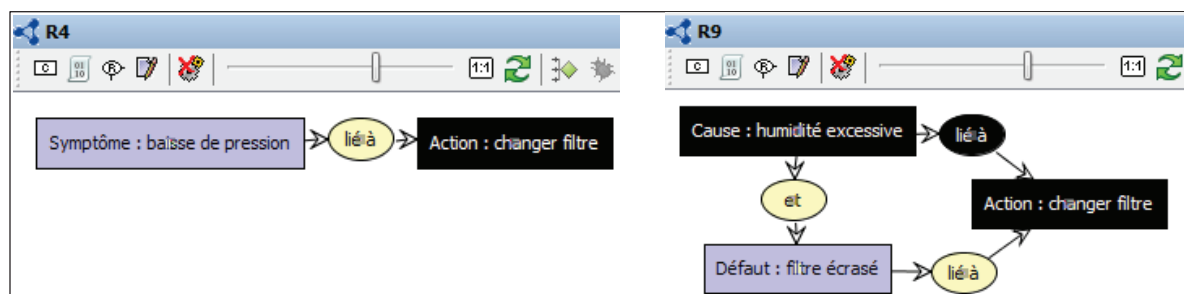


Figure III.12. Illustration de règles avec partie hypothèse inattendue

### Règles avec hypothèse et conclusion inattendues

Finalement, une règle extraite  $R_i \in R$  est inattendue par rapport à  $Q$  si à la fois l'hypothèse et la conclusion de la règle  $R_i$  ne correspondent pas à des projections de  $Q$ . Ces règles inattendues peuvent être pertinentes, mais ne sont pas a priori connues par l'utilisateur ou ne sont pas mentionnées dans ses requêtes.

Parmi l'ensemble de règles extraites, celles qui n'ont pas été retenues dans une des catégories précédentes concernent donc des règles inattendues. Dans notre exemple, il s'agit des règles R3, R5, R6, R8, R10 et R11. Nous présentons dans la Figure III.13 un exemple de règle inattendue par l'utilisateur, i.e. aucune projection de la requête vers les résultats n'a été trouvée.

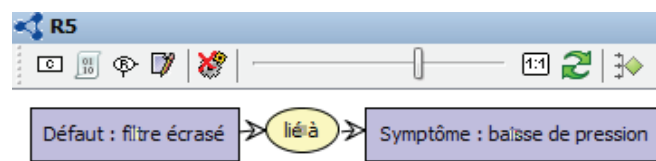


Figure III.13. Illustration de règles inattendues par l'utilisateur

## III.5 BILAN

Ce chapitre a été consacré à l'analyse et à l'évaluation des règles d'association extraites lors de la phase du post-traitement dans le processus d'ECD. Nous avons proposé une approche conceptuelle afin d'interpréter, visualiser et filtrer l'ensemble de règles extraites par l'algorithme d'extraction (ici, l'algorithme *Apriori*). Puisque l'évaluation et la validation des résultats est un aspect critique dans notre approche générale, nous avons suggéré une méthodologie interactive (utilisateur-expert d'ECD et utilisateur-système), en fournissant à l'utilisateur expert du domaine des outils pour rendre les connaissances plus compréhensibles afin qu'il puisse prendre des décisions en fonction de ses intérêts et/ou ses besoins.

Dans un premier temps, nous avons abordé l'importance de la collaboration entre l'expert de *data mining* et l'expert du domaine afin de rendre les processus d'ECD plus pertinents pour les entreprises, surtout lors de la phase de planification (définition des objectifs du processus d'extraction) et de la phase de post-traitement pendant laquelle les résultats sont présentés à l'utilisateur final.

Dans la littérature, plusieurs techniques de post-traitement se sont limitées à la visualisation des résultats, et très peu de travaux de recherche décrivent la façon avec laquelle les règles d'association extraites peuvent être analysées et utilisées de manière concrète (Wu et al., 2003). Cette problématique devient donc un sujet crucial pour les futures recherches en vue d'intégrer des aspects visuels qui facilitent l'exploration des règles extraites, afin de tirer des conclusions et d'évaluer l'utilité des règles dans un domaine d'application donné.

Dans ce contexte, nous avons présenté et illustré une méthodologie originale qui prend en compte trois manières d'évaluer et d'analyser les règles d'association : une évaluation objective afin d'interpréter les mesures d'intérêt associées aux règles, une évaluation sémantique pour analyser l'utilité de chaque règle et visualiser à différents niveaux d'abstraction les relations existantes entre l'ensemble de règles obtenu (relations entre règles et entre *items*), et une évaluation subjective basée sur un mécanisme de requête-réponses en vue de filtrer les règles.

Dans cette phase de post-traitement, les graphes conceptuels ont permis une représentation globale des relations entre les différentes règles extraites (à notre connaissance, il n'existe pas de travaux dans la littérature abordant ce sujet) ainsi que la représentation d'une métarègle synthétisant la connaissance extraite et facilitant l'interprétation des résultats obtenus. De plus, dans le processus de filtrage des règles extraites, nous avons profité des avantages visuels et logiques des GCs et de la connaissance du domaine représentée dans le support ontologique, permettant à l'utilisateur de sélectionner les règles qui lui semblent intéressantes.

Nous verrons dans le chapitre suivant comment cette approche méthodologique peut être instanciée et ajustée sur des cas d'étude réels dans le domaine de la maintenance.

## IV GESTION DES CONNAISSANCES EN MAINTENANCE INDUSTRIELLE

---

IV.1 LA MAINTENANCE INDUSTRIELLE : UN PROCESSUS STRATEGIQUE DANS LES ENTREPRISES ..	91
IV.2 SYSTEMES DE MAINTENANCE BASES SUR LA CONNAISSANCE .....	92
IV.3 CAS D'APPLICATION.....	95
IV.3.1 GENERALITES .....	95
IV.3.2 BASES DE DONNEES CONSIDEREES .....	95
IV.3.3 DU RAPPORT D'INTERVENTION A L'EXPERIENCE FORMALISEE .....	99
IV.3.4 DU RAPPORT D'INTERVENTION A LA CONNAISSANCE .....	104
IV.4 BILAN .....	130

La performance globale des entreprises industrielles dépend largement de leur performance dans le domaine de la maintenance de leurs ressources opérationnelles (Rosqvist et al., 2009). Diverses stratégies de maintenance ont été développées et appliquées à différents domaines industriels, telles que la Maintenance Préventive (Nakagawa, 1986), la Maintenance Prédicative (Huang et al., 2005), la Maintenance Basée sur la Fiabilité (*Reliability-Centered Maintenance* - RCM) (Moubray, 1991) ou la Maintenance Productive Totale (*Total Productive Maintenance* - TPM) (Nakajima, 1988). Néanmoins, l'idée d'une stratégie de maintenance individualisée, dédiée à une entreprise donnée, a émergé avec l'importance croissante accordée au concept d' « entreprise basée sur la connaissance » (Cooke & Leydesdorff, 2006) (Staples et al., 2001).

Ainsi, une gestion appropriée des processus de maintenance est indispensable et la réutilisation efficace des expériences passées et des connaissances acquises apparaît de plus en plus comme un facteur essentiel afin d'améliorer les performances de ces processus. L'application de notre travail s'inscrit dans ce contexte industriel qui nous paraît intéressant pour valider notre approche générale REx-ECD.

La maintenance est par nature un domaine dans lequel une grande masse de données est collectée quotidiennement (Benomrane et al., 2013). Cela est principalement dû à la généralisation d'outils de GMAO (Gestion de Maintenance Assistée par Ordinateur) dans les entreprises. La GMAO rend disponible un volume important d'informations, fournies généralement par des techniciens/opérateurs après une intervention, incluant par exemple l'ordre de travail, la date de l'intervention, la description du problème ou du contexte, la machine concernée, le nom du technicien, le type de défaillance, la localisation fonctionnelle de l'intervention, la cause, les actions réalisées, la durée, etc. Ces informations sont souvent seulement utilisées à des fins de traçabilité et pour le calcul des indicateurs de performances classiques des équipements. En effet, ce registre des interventions passées est stocké dans les entreprises, mais n'est pas nécessairement réutilisé et/ou traité pour analyser ce qui s'est passé et ainsi améliorer les processus.

Des études portant sur la génération de connaissance ont néanmoins déjà été menées dans le domaine de la maintenance : on peut citer par exemple les travaux de Meseroll (Meseroll et al., 2007), Rasovska (Rasovska et al., 2008) et Young (Young et al., 2010). Ces études présentent des résultats intéressants, mais comme nous le verrons par la suite, l'utilisation des méthodes suggérées exige parfois un niveau d'expertise assez élevé. D'autre part, ces méthodes ne visent pas les mêmes objectifs que les nôtres, qui s'inscrivent dans une démarche de retour d'expérience visant à générer des connaissances plus générales à partir d'informations décrivant les interventions passées.

L'objectif de ce chapitre est de démontrer l'applicabilité des principes et techniques présentés dans les chapitres précédents au domaine de la maintenance.

Dans un premier temps, nous présentons le contexte d'étude. Nous décrivons par la suite les différents travaux effectués sur la gestion des connaissances dans ce domaine. Finalement, nous décrivons les cas d'application que nous allons analyser suivant la démarche générale REx-ECD proposée dans ce mémoire. Concernant les types de REx abordés dans la section 1.2.4, notre application se concentre sur un REx basé sur les rapports sur les interventions de maintenance préventive (que l'on peut considérer comme un REx positif), un REx basé sur les rapports sur les interventions de maintenance corrective (que l'on peut considérer comme un REx négatif), un REx cognitif (proposant une meilleure structuration et formalisation des expériences passées afin de faciliter leur partage et leur réutilisation pour la résolution de futurs problèmes) et un REx statistique (proposant une analyse statistique des informations disponibles issues des expériences recueillies afin d'extraire de nouvelles connaissances).

Plus particulièrement, notre approche REx-ECD s'intéresse à la formalisation des connaissances du domaine et des informations caractérisant les interventions passées, et à l'extraction de règles d'association à partir de ces informations, dans l'objectif de générer des connaissances visant à améliorer les futures interventions, et plus généralement les performances de l'entreprise.



## IV.1 LA MAINTENANCE INDUSTRIELLE : UN PROCESSUS STRATEGIQUE DANS LES ENTREPRISES

Les activités de maintenance concernent des actions techniques impliquant des ressources matérielles ainsi que des ressources immatérielles (Rachidi et al., 2013) et nécessitent des prises de décision complexes. Nous nous intéressons ici aux ressources immatérielles, et plus particulièrement à la gestion des connaissances issues de l'expérience afin de supporter la prise de décision.

La maintenance industrielle est une fonction importante qui sert et soutient les processus primaires des organisations (Alsyof, 2009). Actuellement, les entreprises s'investissent de plus en plus dans des activités de gestion des connaissances afin d'améliorer leurs processus de maintenance, et plus généralement leurs performances.

Selon la norme AFNOR NF EN 13306, les activités de maintenance concernent la combinaison de toutes les actions techniques, administratives et de gestion effectuées durant le cycle de vie d'un bien, et ont pour but de « *le maintenir ou de le rétablir dans un état ou dans des conditions données de sûreté de fonctionnement, pour accomplir une fonction requise* » (EN13306:2001, 2001).

Deux types d'actions principaux peuvent être distingués dans ce domaine : celles visant à conserver un service et celles visant à le restaurer. Une taxonomie classique présentée par la norme AFNOR NF EN 13306 distingue donc la maintenance « *préventive* » de la maintenance « *corrective* » (Figure IV.1).

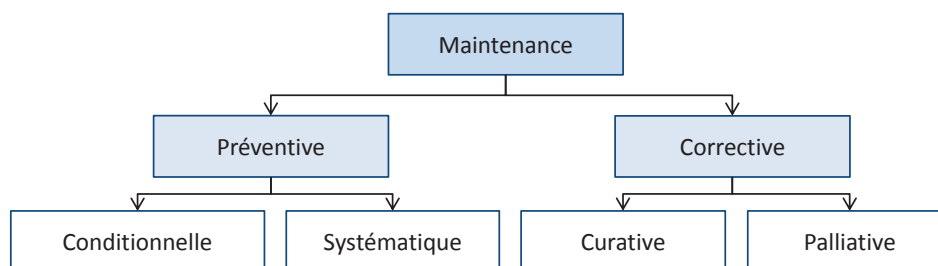


Figure IV.1. Types de maintenance selon la norme AFNOR NF EN 13306

Cette classification dépend des positions respectives dans le temps de la défaillance et de l'activité de maintenance effectuée. Une intervention avant la défaillance est une maintenance préventive, réalisée selon des critères prescrits ou à des dates prédéterminées, et destinée à réduire la probabilité de défaillance. Elle peut être effectuée à des intervalles de temps préétablis ou en fonction d'un nombre défini d'unités d'usage, indépendamment de l'état du bien (*maintenance systématique*), mais elle peut être aussi basée sur une surveillance du fonctionnement du bien et/ou des paramètres significatifs de ce fonctionnement (*maintenance conditionnelle*). Une intervention après la détection d'une défaillance est une activité de maintenance corrective, destinée à remettre un bien dans un état dans lequel il peut accomplir une fonction requise. Cette intervention peut être

à caractère provisoire (*maintenance palliative*) ou définitif et immédiat (*maintenance curative*) (EN13306:2001, 2001).

Aujourd'hui, la performance des systèmes de production dépend de leur performance en maintenance. Toutefois, la complexité croissante et le niveau d'automatisation des équipements industriels ne facilitent guère le travail des acteurs qui ont en charge de les manipuler, de les diagnostiquer et de les maintenir efficacement. Les tâches de maintenance deviennent en effet de plus en plus complexes et diversifiées, et concernent non seulement des activités sur les systèmes mécaniques, mais aussi sur les systèmes électromécaniques, hydrauliques, électroniques et logiciels (Alsyouf, 2009). Opérateurs, mainteneurs et experts en maintenance considèrent ainsi souvent que le principal obstacle à la mise en œuvre de processus efficaces est le manque de connaissances sur le système industriel et sur les processus opérationnels (Crespo Marquez & Gupta, 2006). Les connaissances des différents acteurs en maintenance ne pouvant pas couvrir tous ces domaines, nous soulignons ici l'importance de fournir aux acteurs une aide à la décision afin d'améliorer les plans de maintenance actuels et de mieux définir les mesures de prévention à mettre en œuvre, ainsi que les procédures et les actions menées lors des futures interventions.

Pour situer notre apport dans ce domaine, nous présentons dans la suite un panorama d'études visant à la construction de systèmes à base de connaissances en maintenance, présentant des résultats intéressants en ce qui concerne le REx et/ou l'ECD.

## IV.2 SYSTEMES DE MAINTENANCE BASES SUR LA CONNAISSANCE

Les connaissances acquises lors d'une intervention de maintenance passée peuvent être réutilisées pour améliorer les prochaines interventions (Ben-Daya et al., 2009). Dans ce contexte et compte tenu de la complexité de mise en œuvre de stratégies comme la TPM (*Total Productive Maintenance*) et la RCM (*Reliability-Centered Maintenance*), les organisations s'orientent de plus en plus vers des solutions utilisant leurs connaissances internes (Hogan et al., 2011).

Plusieurs systèmes à base de connaissances ont été développés et améliorés par des méthodes de gestion des connaissances/expériences ou des méthodes d'extraction de connaissances, et ont permis de définir une mémoire d'entreprise aidant les décisions de maintenance. Les principaux objectifs de ces systèmes incluent notamment la programmation des tâches, le diagnostic des machines ou encore le choix d'une stratégie de maintenance.

En ce qui concerne la **gestion des connaissances/expériences**, une méthode de capitalisation des connaissances en maintenance est par exemple décrite dans (Rasovska et al., 2008), qui vise à développer un système d'aide à la décision pour le *diagnostic* et la *réparation* des équipements. Détecter, préserver, capitaliser et actualiser les connaissances stratégiques sont les grandes étapes de cette approche qui est principalement basée sur le Raisonnement à Partir de Cas (RàPC), et qui utilise des techniques de représentation des connaissances qui n'ont pas de mécanismes de raisonnement. MAIC (Materiali per Apparecchiature di Impianti Chimici) (Pieri et al., 2002) est un autre système d'aide à la décision à base de connaissances, qui utilise aussi le RàPC pour la *résolution de problèmes* dans la maintenance d'une usine chimique, tout en prenant en considération des variables économiques dans les processus. Un autre système, appelé EXPERT-MM (Batanov et al.,

1993), suggère des *stratégies* de gestion de maintenance à partir des connaissances expertes stockées dans une base d'informations. Dans (Chassiakos et al., 2005) est décrit un système visant plutôt la *planification* de la maintenance des ouvrages du génie civil (en particulier de ponts) à partir d'une base de connaissances.

D'autres travaux se concentrent surtout sur **l'extraction de connaissances** à partir des données disponibles, à l'aide de différentes techniques de *data mining*. Des outils traditionnels d'analyse de données ont ainsi été utilisés avec succès dans l'amélioration de la qualité des processus de maintenance. De nouveaux outils ont aussi été développés pour mieux exploiter les grandes quantités de données disponibles collectées par les systèmes informatisés dans l'industrie (Köksal et al., 2011), i.e. les techniques d'ECD.

Dans ce contexte, selon Choudhary (Choudhary et al., 2009), seulement 8% des études portant sur l'extraction de connaissances (*data mining*) dans l'industrie manufacturière sont liées au domaine de la maintenance. Cette constatation est cohérente avec le travail présenté dans (Harding et al., 2005) indiquant que même si ce secteur a été le premier domaine manufacturier profitant de solutions basées sur les techniques de *data mining*, seules quelques études ont été identifiées dans des applications de maintenance. Nous présentons dans la suite des exemples de *data mining* prédictif et descriptif dans ce domaine.

Un système d'aide à la décision permettant de *prédire la défaillance* d'un composant est suggéré dans (Létourneau et al., 1999), pour améliorer la maintenance préventive à partir de données recueillies par des capteurs. Dans (Shen et al., 2000), la théorie des ensembles approximatifs (*rough-sets*) est appliquée afin de *diagnostiquer des défauts* à partir de l'extraction de règles d'association décrivant les liens de causalité menant à des défaillances. D'autres modèles visent plutôt à mieux comprendre les données disponibles afin d'aider les processus de prise de décision stratégique dans le domaine. Certaines études suggèrent ainsi que l'extraction de règles d'association permet *d'améliorer les procédures et les interventions* ; d'autres ne ciblent pas explicitement une telle amélioration, mais fournissent des résultats intéressants qui peuvent être utiles à cet effet (le lecteur intéressé pourra par exemple reporter à (Chen et al., 2005) (Liu et al., 2011) (Meseroll et al., 2007) (Mirabadi & Sharifian, 2010)).

Les principales caractéristiques de ces études liées à l'extraction de règles d'association en maintenance, sont résumées dans le Tableau IV.1. Ce tableau liste pour chaque étude présentée l'algorithme utilisé pour l'extraction des règles, l'objectif principal de l'étude, le contexte d'application dans la maintenance et la source des données.

Nous pouvons constater dans le Tableau IV.1 que l'algorithme *Apriori* (Agrawal & Srikant, 1994), bien connu pour sa simplicité et son efficacité, est souvent utilisé pour l'extraction des règles. Nous remarquons également que les objectifs de ces études sont souvent assez précis, ce qui facilite notablement l'interprétation, l'évaluation et la validation des règles obtenues, toutes basées sur les liens entre les mêmes attributs classiques (par exemple, la relation symptômes-causes pour une activité de *data mining* visant au diagnostic). Peu de ces études se sont intéressées à l'obtention des différents types de règles que l'algorithme pourrait identifier à partir de l'analyse des attributs qui caractérisent les interventions dans une base de données.

Tableau IV.1. Revue des études liées à l'extraction de règles d'association en maintenance

Référence	Algorithme/Outil	Objectif/Description	Contexte d'application	Source des données
(Chen et al., 2005)	Apriori/	Trouver les corrélations entre machines et produits défectueux	Fabrication de semi-conducteurs	Système de suivi
(Zhang & Yang, 2006)	Apriori/	Identifier les erreurs humaines pour améliorer la maintenance des aéronefs	Transport aéronautique militaire	"Systèmes d'information"
(Meseroll et al., 2007)	/ThinkAnalytics	Analyser les résultats dans les tests pour identifier les écarts.	Industrie aéronautique	Built-in-tests
(Young et al., 2010)	Apriori/Clementine	Relations entre défaillances, diagnostics et actions de réparation pour améliorer les pratiques de maintenance	Industrie aéronautique	Rapports de maintenance
(Mirabadi & Sharifian, 2010)	GRI (Generalized Rule Induction) /Clementine	Analyser les causes des accidents dans un réseau ferroviaire	Transport ferroviaire	Rapports sur les accidents
(Baohui et al., 2011)	Apriori/	Trouver des associations entre symptômes et actions correctives	Industrie aéronautique	Rapports de maintenance
(Liu et al., 2011)	Algorithme Apriori basé sur SQL/	Trouver des relations entre défaillances dans la surveillance de l'état d'un tunnel ferroviaire	Ferroviaire	Rapports sur les défaillances
(Maquee et al., 2012)	Clustering+Apriori/	Analyser l'efficacité des activités de maintenance par la technique de clustering	Maintenance de bus	Rapports de maintenance
(Sammouri et al., 2012)	T-patterns/	Associations temporelles entre les alarmes et l'occurrence de défaillances sévères	Systèmes de transport ferroviaire	Système de données de type FCD ( <i>Floating Car Data</i> )
(Kamsu-Foguem et al., 2013)	FP-Growth/	Règles séquentielles pour améliorer les processus de production (incluant la maintenance)	Manufacturière	Système de suivi

Il ressort de ces travaux un besoin en systèmes d'aide à la décision exploitant les connaissances issues de l'analyse des expériences passées (Chapitre II). L'idée de fournir à l'utilisateur une aide pour comprendre et interpréter les nouvelles connaissances découvertes et les liens possibles entre les

différents types de règles extraites par un algorithme devient alors d'un grand intérêt (Chapitre III). Un autre point intéressant concerne le fait que l'interaction avec les utilisateurs experts du domaine n'est pas détaillée dans ces travaux, sauf dans (Mirabadi & Sharifian, 2010) où l'utilisation de l'algorithme « *Generalized Rule Induction* » permet à l'utilisateur de spécifier le nombre de règles qu'il attend.

## IV.3 CAS D'APPLICATION

### IV.3.1 GENERALITES

Dans nos différents cas d'application, nous nous intéressons principalement aux interventions techniques ayant eu lieu sur les équipements des processus de production, enregistrées par les opérateurs de maintenance dans des bases de données de logiciels de Gestion de Maintenance Assistée par Ordinateur (GMAO).

Une GMAO comporte plusieurs modules et permet de programmer et de suivre les activités de maintenance dans une entreprise, tout en gérant l'information relative aux ressources humaines, aux pièces détachées, aux plannings d'interventions et aux historiques des équipements (Swanson, 2003) selon les besoins spécifiques de chaque entreprise. Parmi les différents modules d'une GMAO, nous sommes particulièrement intéressés par le module de la gestion des équipements (permettant de localiser les interventions), et le module de gestion de maintenance (fournissant un historique et une description détaillée des interventions effectuées).

Nous avons analysé trois bases de données réelles mais anonymées portant sur des rapports d'opérations de maintenance (historique du bien<sup>7</sup>) effectuées dans des entreprises des secteurs aéronautique et pharmaceutique. Nous allons présenter et analyser quelques résultats issus des trois entreprises considérées.

### IV.3.2 BASES DE DONNEES CONSIDEREES

Pour illustrer notre démarche REx-ECD, nous partons des historiques des biens de chaque entreprise, portant sur des volumes très différents de transactions (i.e. interventions de maintenance) et d'attributs relatifs à ces interventions.

Ces historiques ont tous été extraits du module de gestion de maintenance de l'ERP SAP ECC, leader des ERP (*Enterprise Resource Planning*), systèmes d'informations les plus répandus dans les

---

<sup>7</sup> Historique du bien : tous les événements, qu'ils soient de maintenance préventive ou corrective, liés à des mises en service, des arrêts, des travaux, etc., sont consignés pour constituer l'historique du bien et en permettre la traçabilité.

entreprises. Après export des bases de données de SAP vers Excel, nous avons analysé les rapports portant sur les interventions de maintenance (correspondant aux *transactions* dans le contexte de l'extraction de règles d'association) ayant divers champs ou attributs (chaque attribut et sa valeur correspondant à un *item*).

Notre démarche complète a été appliquée sur ces historiques des biens : d'abord la formalisation du support ontologique, puis la formalisation des expériences à l'aide des GCs (pour chacune des interventions) ; enfin l'extraction de règles d'association à partir de cette information.

Pour cela, nous devons tout d'abord *sélectionner* l'information pertinente (i.e. les interventions et les attributs caractérisant ces interventions) pour être traitée et *nettoyer* les données afin d'homogénéiser la base de données de départ en corrigeant les fautes de frappe et en retrouvant les valeurs manquantes, avant de passer à la phase de traitement et d'extraction de la démarche. Il s'agit d'une phase qui demande un travail important et une certaine expertise du domaine. Pour la réaliser, nous avons eu besoin de la collaboration du responsable maintenance de chaque entreprise, afin de mieux comprendre le contexte d'application et de cibler les données les plus pertinentes.

Les principales caractéristiques des feuilles Excel obtenues après la phase de nettoyage sont présentées ci-dessous.

#### IV.3.2.1 Cas # 1

Le premier cas comporte **1067** interventions de maintenance effectuées sur des ponts roulants entre mars 2010 et mars 2012. Ces équipements permettent le levage et le transfert de charges lourdes et sont utilisés ici dans le secteur aéronautique.

Après les étapes de sélection et de nettoyage des données, nous avons conservé 13 attributs qui sont : le numéro de l'ordre de travail, le type d'équipement, la caractéristique de l'équipement, la zone d'emplacement, la zone de défaut, la cause de l'intervention, l'action de maintenance effectuée (solution), le type d'intervention, le poste affecté, le type de pointage, la date de l'intervention, la durée, le nom des techniciens (masqué pour des raisons de confidentialité).

Nous présentons sur la Figure IV.2 certains des attributs pris en compte dans l'analyse.

#### IV.3.2.1 Cas # 2

Le deuxième cas considère **5955** rapports d'interventions techniques réalisées sur différents types d'équipements dans une autre entreprise du secteur aéronautique entre janvier 2010 et août 2013. Nous avons initialement retenu 19 attributs : la date de saisie de l'ordre de travail, le numéro de l'ordre de travail, la fréquence, le libellé, l'état, la production, la nature, la priorité, le type d'équipement, la famille, le matricule, le modèle, l'équipe, la technologie, le temps prévu, le temps réalisé, le pourcentage de réalisation, la date de fin saisie, la section analytique.

Nous présentons sur la Figure IV.3 le type d'information considérée et certains des attributs pris en compte.



A	B	D	F	H	J	L	N
N° OT	Equipement	Caractéristique	Zone d'emplacement	Zone de défaut	Cause	Action effectuée	Type d'intervention
698188	Pont 2	2 crochets	C02	Translation	Assistance	Recalage	Assistance Technique (Et
698221	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
698457	Pont 6	2 crochets	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
698555	Pont 2	2 crochets	C02	Levage	Carte électronique	Amélioration fonctionnement	Amélioration (TA)
698569	Pont 3	2 crochets	C02	Levage	Carte électronique	Amélioration fonctionnement	Amélioration (TA)
698782	Pont 6	2 crochets	C03	Télécommande	Pupitre de commande	Information /sensibilisation	Correctif Urgent (CU)
699249	Pont 2	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
699253	Pont 5	1 crochet	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
699263	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
699283	Pont 2	2 crochets	C02	Télécommande	Arrêt d'urgence enclenche	Re enclenchement – acquitter	Correctif Urgent (CU)
699565	Pont 2	2 crochets	C02	Translation	Frein	Nettoyage	Correctif Différé (CD)
699569	Pont 6	2 crochets	C03	Levage	Relayage	Re enclenchement – acquitter	Correctif Urgent (CU)
699911	Pont 8	1 crochet	C04	Télécommande	Arrêt d'urgence enclenche	Re enclenchement – acquitter	Correctif Urgent (CU)
699925	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
699926	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
700147	Pont 6	2 crochets	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
701175	Pont 2	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
702038	Pont 5	1 crochet	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
702502	Pont 2	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
702503	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Correctif Différé (CD)
702507	Pont 8	1 crochet	C04	Télécommande	Boitier tlce casse	Remplacement pièce	Correctif Urgent (CU)
702508	Pont 6	2 crochets	C03	Direction	Variateur de vitesse	Re enclenchement – acquitter	Correctif Urgent (CU)
703325	Pont 3	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
703373	Pont 6	2 crochets	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
703702	Pont 2	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
703885	Pont 8	1 crochet	C04	Télécommande	Arrêt d'urgence enclenche	Re enclenchement – acquitter	Correctif Urgent (CU)
704062	Pont 2	2 crochets	C02	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
704072	Pont 5	1 crochet	C03	Translation	Défaut angulaire	Recalage	Assistance Technique (Et
704399	Pont 2	2 crochets	C02	Télécommande	Carte électronique	Réparation pièce	Correctif Différé (CD)
704544	Pont 6	2 crochets	C03	Translation	Variateur de vitesse	Modification	Correctif Différé (CD)

Figure IV.2. Rapports des interventions de maintenance (Cas # 1)

A	C	E	F	G	H	I	J	L
Date de saisi	# OT	Etat	Production	Nature	Priorité	Type d'équipement	Famille	Modèle
07-janv-10	137477-01	Soldé	Site2	CSP	Sans urgence	X100		CHARIOT_PEINTURE
07-janv-10	137478-01	Soldé	Site1	CSP	Sans urgence	X50/X60		CHARIOT_CAPOT
07-janv-10	137479-01	Soldé	Site1	CSP	Sans urgence	X50		CHARIOT_CAPOT
11-janv-10	137303-01	Soldé	Site2	COR	Sans urgence	X100		CHARIOT_PEINTURE
01-févr-10	130439-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
03-févr-10	128059-01	Soldé	Site2	COR	Sans urgence	X50		CHARIOT_PEINTURE
14-févr-10	138954-01	Soldé	Site2	CSP	Sans urgence	PEINTURE	CABINES	CABINES
14-févr-10	138955-01	Soldé	Site2	CSP	Sans urgence	PEINTURE	CABINES	CABINES
15-févr-10	138956-01	Soldé	Site2	PRV_COND	Sans urgence	PEINTURE	CABINES	CABINES
16-févr-10	138957-01	Soldé	Site2	COR	Sans urgence	X100		CHARIOT_PEINTURE
18-févr-10	138958-01	Soldé	Site2	COR	Sans urgence	X100		CHARIOT_PEINTURE
25-févr-10	130576-01	Soldé	Site2	COR	Sans urgence	PEINTURE	CABINES	CABINES
05-mars-10	128964-01	Soldé	Site3	CSP	Sans urgence	FM_ADMIN_1_SA	FM TOULOUSE	FM_PLOMBE_SANITAIRE
09-mars-10	129012-01	Soldé	Site1	COR	Sans urgence	Y7X		CHARIOT_PEINTURE
11-mars-10	139921-01	Soldé	Site2	COR	Sans urgence	PEINTURE	CABINES	CABINES
12-avr-10	129971-01	Soldé	Site12	COR	Sans urgence	X100		FRAME
14-avr-10	130531-01	Soldé	Site2	PRV_COND	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129911-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129912-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129913-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129914-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129915-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129916-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
18-avr-10	129917-01	Soldé	Site2	PRV	Sans urgence	PEINTURE	CABINES	CABINES
19-avr-10	130227-01	Soldé	Site2	COR	Sans urgence	PEINTURE	CABINES	CABINES

Figure IV.3. Rapports des interventions de maintenance (Cas # 2)

Ce deuxième cas a nécessité de nombreuses explications/éclaircissements portant sur les attributs retenus afin de mieux interpréter les résultats des extractions. Par exemple, nous avons discuté avec le responsable de maintenance le lien entre « type d'équipement », « famille », « matricule » et « modèle » (lien non détaillé ici). L'attribut « production » décrit ici le site dans lequel sont réalisées

les interventions. La « priorité » des interventions est liée à l'urgence de l'intervention elle-même et pas à l'équipement considéré.

#### IV.3.2.2 Cas # 3

La troisième base de données concerne **537** rapports de maintenance issus d'opérations de maintenance effectuées entre janvier 2012 et janvier 2013 dans une entreprise du secteur pharmaceutique.

Chaque intervention est caractérisée par 17 attributs : le numéro de l'ordre de travail, la date de création, le type d'ordre, la nature de l'objet (sur lequel est survenu le problème), l'effet de la défaillance, le type de cause, le groupe de planification, la criticité de l'équipement, le site, l'unité, le groupe d'équipement, le nom de l'équipement, le nom de l'appareil, l'opérateur ayant saisi l'ordre de travail, le coût réel de l'opération, la durée de la défaillance, la description de l'objet.

Nous présentons sur la Figure IV.4 certains des attributs pris en compte dans l'analyse.

A	B	C	E	G	I	L	P
Order	Created on	Order Type	Object part Descript	Damage Description	Cause code descripti	Planner group	Asset criticality
50070976	27/01/2012	PM11	0C: Temperature Device	0C: General failure	0C: Component failure	PU1	#N/A
50070978	27/01/2012	PM11	0C: Pump	0C: Mechanical damage	0C: Component failure	PU1	#N/A
50071070	28/01/2012	PM11	0C: Piping & fittings	0C: Leak	0C: Leak	PU1	High
50071071	28/01/2012	PM11	0C: MCR	0C: Impact damage	0C: Component failure	PU1	Medium
50071072	28/01/2012	PM11	0C: Lighting	0C: Mechanical damage	0C: Component failure	PU1	Medium
50071073	28/01/2012	PM11	0C: Level Device	0C: Instrument failure	0C: Instrument failure	PU1	High
50071074	28/01/2012	PM11	0C: Analytical device	0C: Instrument failure	0C: Communication fault	PU1	High
50071075	29/01/2012	PM11	0C: Piping & fittings	0C: Leak	0C: Wear and tear	PU1	Medium
50071140	30/01/2012	PM11	0C: Flow Device	0C: Integrity test failed	0C: No mechanical fault identified	PU1	High
50071209	01/02/2012	PM11	0C: Heat Exchanger	0C: Leak	0C: Component failure	PU1	Medium
50071217	01/02/2012	PM11	0C: Pressure Device	0C: Physical damage	0C: Impact damage	PU1	#N/A
50071231	02/02/2012	PM11	0C: Piping & fittings	0C: Physical damage	0C: Instrument failure	PU1	#N/A
50071233	02/02/2012	PM11				PU1	#N/A
50071234	03/02/2012	PM11	0C: Pump	0C: Leak	0C: Wear and tear	PU1	Medium
50071244	03/02/2012	PM11	0C: Piping & fittings	0C: Leak	0C: Joint failure	PU1	Medium
50071250	05/02/2012	PM11	0C: Small Power	0C: General failure	0C: None of the above	PU1	Medium
50071251	05/02/2012	PM11	0C: Piping & fittings	0C: Leak	0C: Leak	PU1	High
50071295	06/02/2012	PM11				PU1	Low
50071298	06/02/2012	PM11	0C: Heat Exchanger	0C: Leak		PU1	Medium
50071300	07/02/2012	PM11	0C: Pressure Device	0C: Zero / span Error	0C: None of the above	PU1	Medium
50071320	08/02/2012	PM11	0C: Piping & fittings	0C: Insufficient Flow	0C: No mechanical fault identified	PU1	High
50071333	08/02/2012	PM11	0C: Pressure Device	0C: Physical damage	0C: No mechanical fault identified	PU1	Medium
50071343	09/02/2012	PM11	0C: Position	0C: Position not made	0C: Untightened	PU1	High
50071345	10/02/2012	PM11	0C: Temperature Device	0C: Position not made	0C: No instrumentation fault identif	PU1	High
50071358	10/02/2012	PM11	0C: Piping & fittings	0C: Leak	0C: Wear and tear	PU1	Medium

Figure IV.4. Rapports des interventions de maintenance (Cas # 3)

Nous allons présenter dans la suite de ce chapitre la phase de traitement du REx associée à la phase de *data mining* dans la démarche REx-ECD, i.e. les étapes pour la structuration de la *base d'expériences* (IV.3.3) et les étapes pour la génération de la *base de règles* (IV.3.4).



### IV.3.3 DU RAPPORT D'INTERVENTION A L'EXPERIENCE FORMALISEE

Pour la formalisation du vocabulaire du domaine et des informations caractérisant les expériences passées, plusieurs plates-formes et outils de mise en œuvre des GCs ont été proposées dans la littérature (Baget et al., 2008), permettant de définir un support ontologique (*S*) et de construire les GCs associés.

Nous avons choisi la plate-forme CoGui pour cette mise en œuvre : CoGui editor<sup>8</sup> (Figure IV.5), développé en langage Java, est un outil libre basé sur les graphes qui permet de construire et de manipuler des structures visuelles intuitives avec des capacités de raisonnement par projection (par exemple des requêtes sur les graphes). Essentiellement, la plateforme CoGui permet de construire un support (*TC*, *TR*) et un ensemble de GCs représentant les assertions, souvent appelées « faits ». Dans notre contexte REx-ECD, ces « faits » sont dénommés « expériences » (stockées dans la *base d'expériences*), et « règles » (stockées dans la *base de règles*).

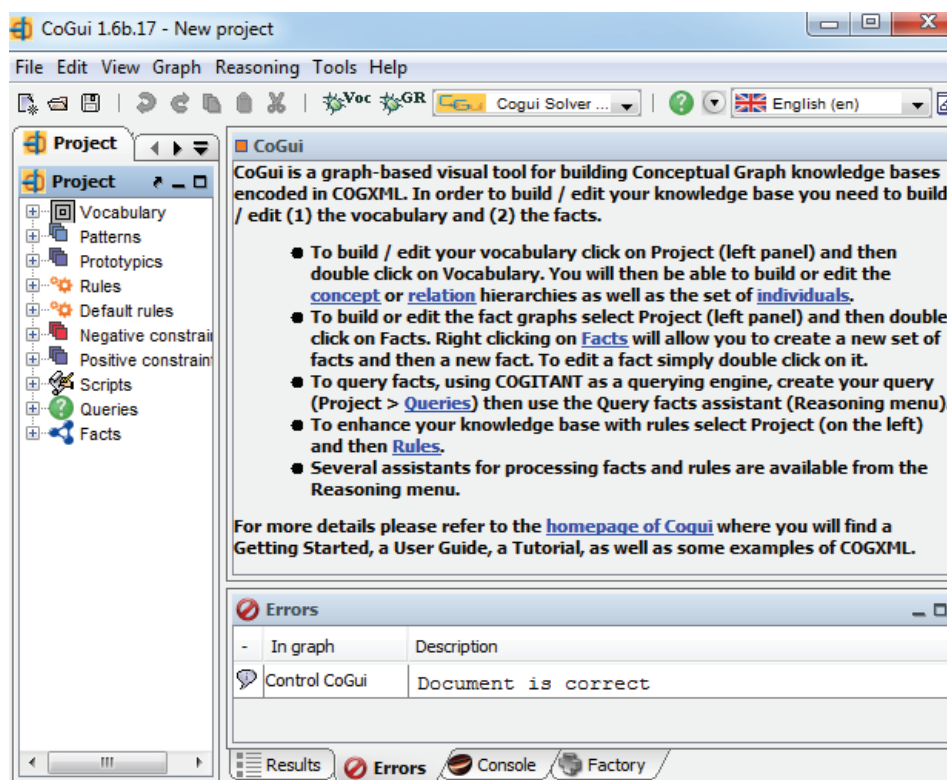


Figure IV.5. Interface de l'outil CoGui editor

<sup>8</sup> <http://www2.lirmm.fr/cogui/>

### IV.3.3.1 Définition du support dans les GCs

Les composants d'une ontologie au niveau conceptuel (vocabulaire du domaine) sont représentés à l'aide de la plateforme CoGui dans un modèle général simplifié pour chaque cas d'application considéré dans le domaine de la maintenance.

Pour le cas d'application # 1, nous illustrons sur la Figure IV.6 et la Figure IV.7 la hiérarchie réduite des types de concepts (*TC*) et la hiérarchie des types de relations (*TR*) respectivement, décrivant les informations issues des interventions de maintenance sur les ponts roulants. Chacune de ces hiérarchies, présentées dans deux interfaces différentes, permet de construire et de modifier les graphes dans CoGui.

Les hiérarchies *TC* et *TR* fournissent donc le support (*S*), qui est la base de la représentation des connaissances dans la base de REX-ECD. Dans nos cas d'application, *S* permet essentiellement de modéliser les équipements et de contextualiser et décrire d'une façon plus intuitive les interventions de maintenance, tout en tenant compte des trois principaux éléments clés qui définissent une expérience (section II.3.2.1) : le contexte, l'analyse et la solution.

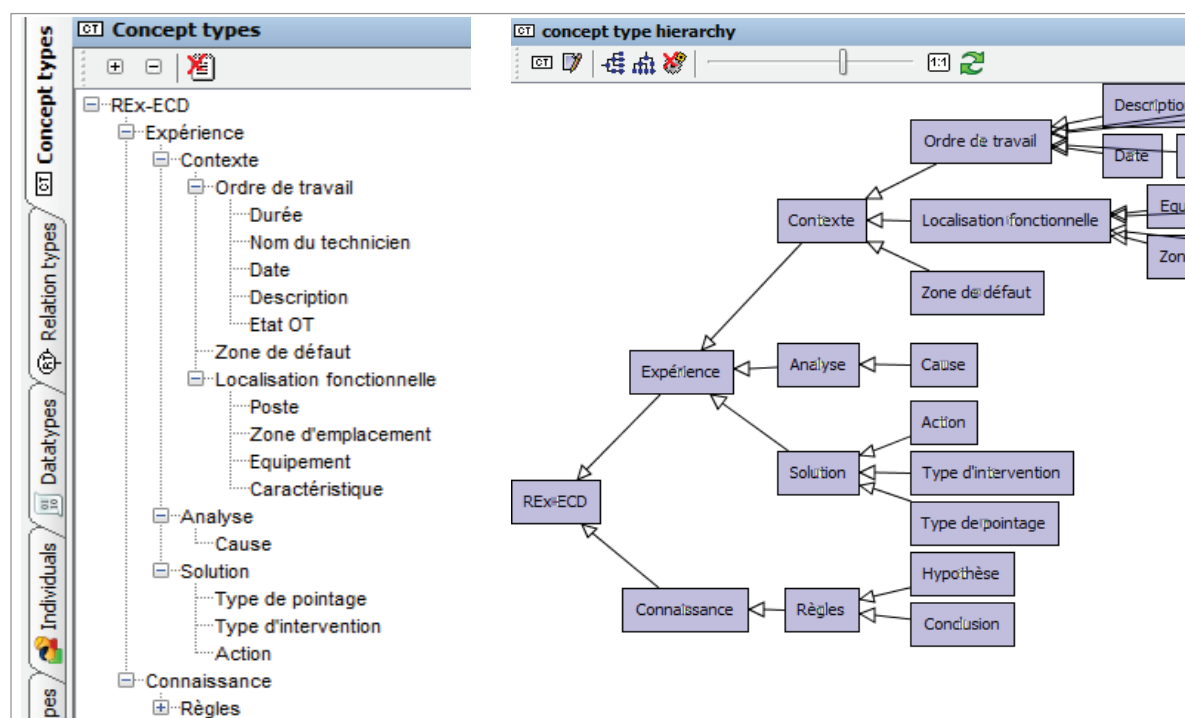


Figure IV.6. Visualisation d'une partie de la hiérarchie des types de concepts (*TC*) modélisées dans CoGui (Cas # 1)

Pour la représentation des expériences à partir de *TC* (Figure IV.6), nous avons défini la partie « *contexte* » qui décrit la situation générale dans laquelle s'est produit l'événement déclencheur sur le pont roulant (incluant par exemple l'*ordre de travail*, la *localisation fonctionnelle* de l'équipement concerné, la *zone de défaut*, le *technicien* concerné) ; la partie « *analyse* » expose la cause principale du problème ou l'intervention ; finalement, la « *solution* » décrit principalement le *type*

d'intervention associé et les *actions* qui ont été effectuées pour résoudre le problème spécifique (i.e. les activités de maintenance menées).

Contrairement à *TC*, *TR* exprime les relations fondamentales des ontologies qui seront utilisées pour représenter les GCs dans les trois cas d'application. Dans la Figure IV.7, nous représentons des relations génériques comme la relation « temporelle » (i.e. *avant*, *après*, *en parallèle*), la relation « spatiale » (i.e. *dans*, *dehors*), la relation « logique » (i.e. *implique*, *et*), la relation "usuelle" (i.e. *objet*, *attribut*, *agent*, *concerne*, etc.) (Breuker, 2013), ainsi que d'autres relations spécifiques à notre étude comme la « relation d'expérience » (i.e. *nécessite*, *génère*) ou « élément de ». Ces relations permettent donc de relier les différents types de concepts liés à la représentation des expériences passées et des nouvelles connaissances extraites.

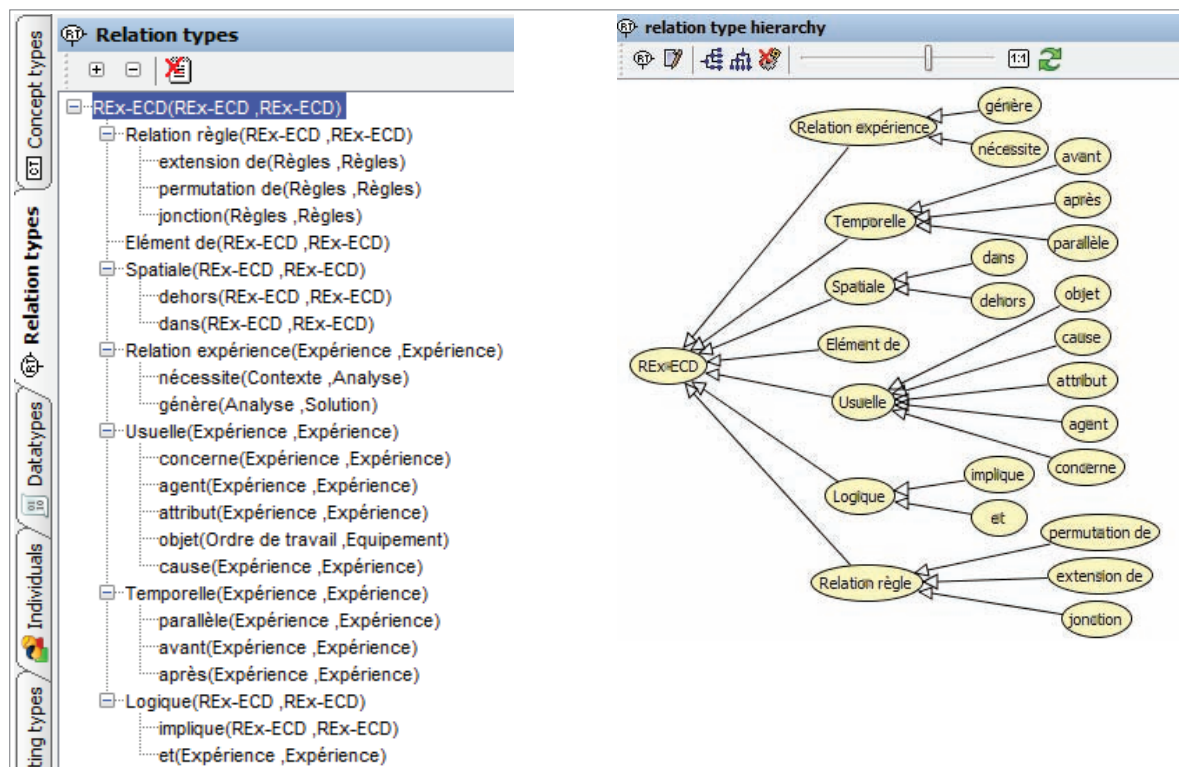


Figure IV.7. Hiérarchie des types de relation (TR) modélisées dans CoGui

### IV.3.3.2 Formalisation des expériences

#### Modèle générique d'une expérience

A partir d'un support (*S*) défini pour chaque cas d'application, nous avons proposé un graphe conceptuel du modèle générique pour représenter chaque expérience à partir des informations disponibles dans chaque cas.

Pour le cas # 1, nous proposons sur la Figure IV.8 un modèle structuré permettant de représenter les expériences, incluant les attributs les plus significatifs sélectionnés pour décrire une intervention.

Ce modèle doit être adapté aux autres cas d'application, notamment en fonction des caractéristiques des informations disponibles et des restrictions de chaque contexte d'application.

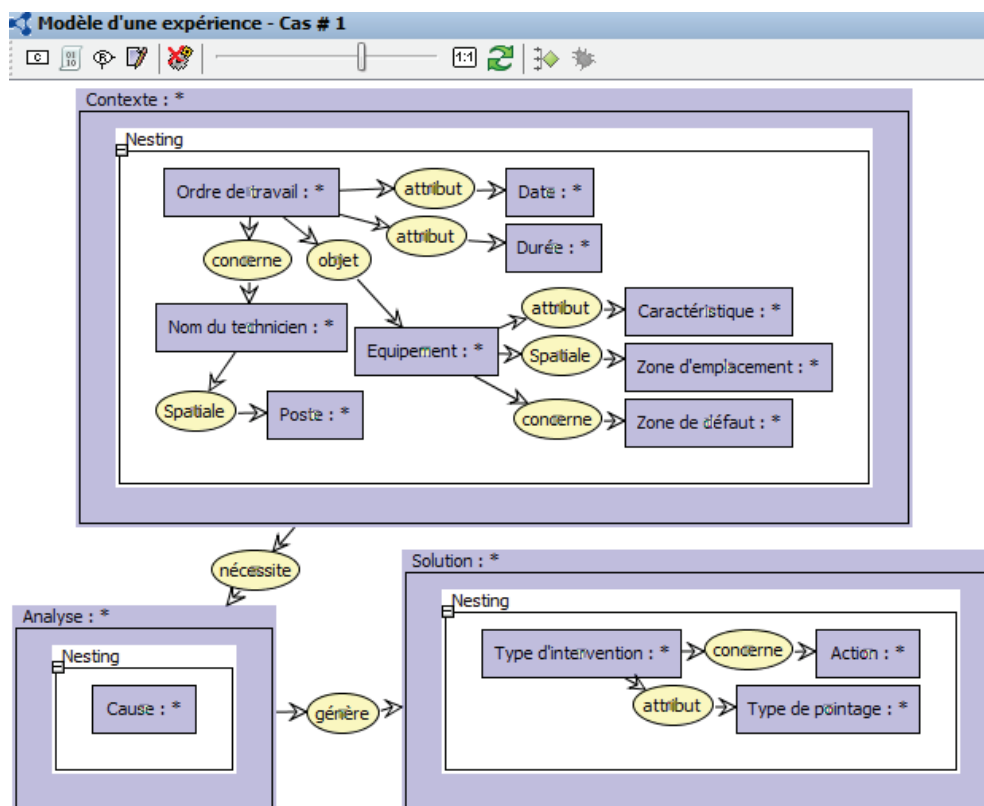


Figure IV.8. Modèle générique d'une expérience (Cas # 1)

Le graphe conceptuel de la figure peut se lire de la manière suivante : une expérience concerne un *contexte*, une *analyse* et une *solution* (les trois sous-graphes de la Figure IV.8). Le *contexte* est décrit par un *ordre de travail* concernant un *technicien* et un *poste de travail*. Cet ordre a comme objet un *équipement* (avec certaines *caractéristiques* spécifiques et localisé dans une *zone* déterminée) qui a eu une *défaillance*. Cette situation nécessite une *analyse* afin de trouver la *cause* principale du problème. Une fois la cause déterminée, une *solution* va être générée incluant le *type d'intervention* à effectuer, les *actions* correspondantes et le *type de pointage*.

### Formalisation des expériences à l'aide des GCs

Une fois le modèle générique évalué et validé par l'expert du domaine, nous passons à la représentation de chacune des expériences contenues dans le fichier Excel. Poursuivant le premier cas application, nous prenons comme exemple ici les événements survenus sur les ponts roulants, qui sont la base de chaque expérience dans ce contexte.

En fonction du modèle générique prédéfini et du support ( $S$ ), le GC de la Figure IV.9 est construit, représentant les informations portant sur une intervention réalisée. Le graphe de la Figure IV.9 peut être interprété de la façon suivante : dans l'expérience  $E_1$ , le contexte  $C_1$  nécessite une analyse  $A_1$ , ce qui a généré la solution  $S_1$ .

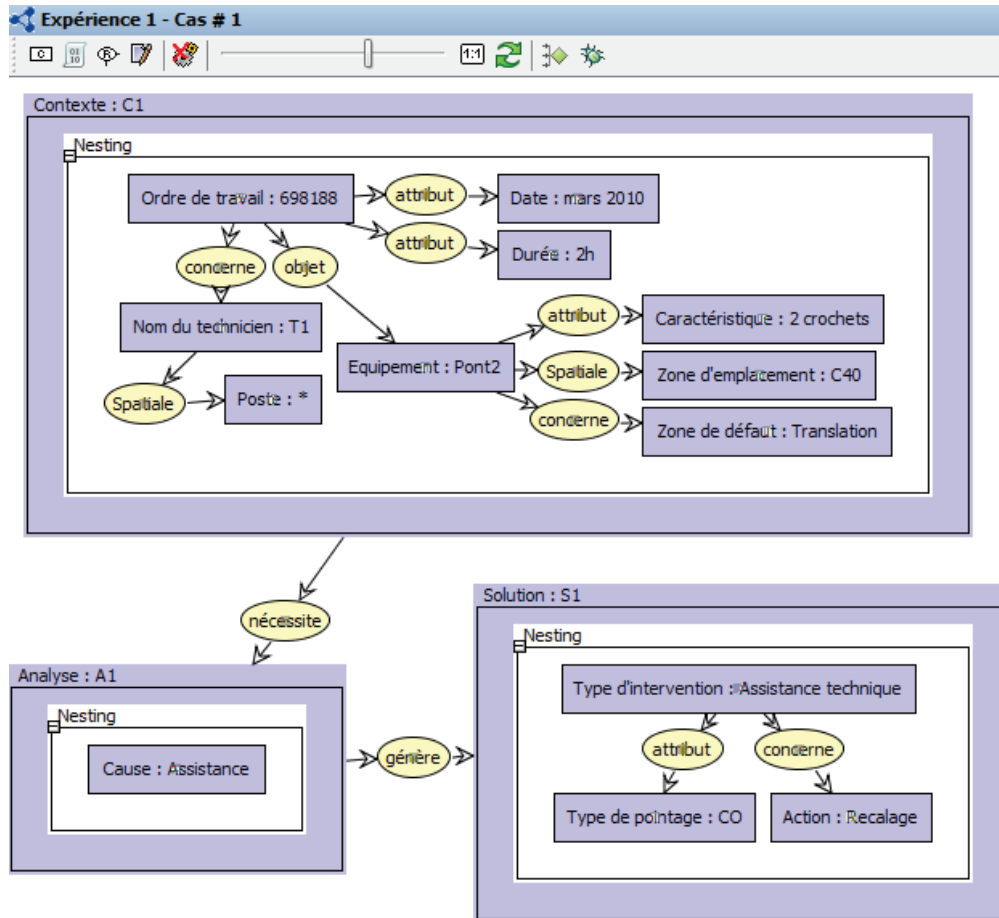


Figure IV.9. Illustration d'une intervention du cas #1

Par exemple, le *contexte*  $C_1$  est décrit par l'ordre de travail n° 698188 pour l'équipement Pont2. Nous distinguons ici l'objet de l'action de maintenance, la zone de défaut (le mouvement de translation) et des données supplémentaires utilisées pour mieux décrire le contexte de l'événement (nom du technicien, poste, date, caractéristique de l'équipement, ...). Pour la description de l'*analyse*  $A_1$ , nous décrivons la cause principale de l'intervention (dans ce cas, il s'agissait d'une demande d'assistance pour l'équipement). Finalement, la description de la *solution* concerne le type d'intervention effectuée, le type de pointage et les actions réalisées (ici, une assistance technique de type « correctif - CO » consistant en un réalignement de l'instrument). La durée de cette intervention a été de 2 heures.

Ces expériences formalisées sont ensuite stockées dans la base d'expériences contenue dans la base de REx-ECD pour une réutilisation future.

### IV.3.4 DU RAPPORT D'INTERVENTION A LA CONNAISSANCE

#### IV.3.4.1 Préparation des données

A partir de l'information semi-structurée extraite des bases de données de maintenance, nous avons d'abord, à l'aide de l'expert du domaine, sélectionné et nettoyé les données d'intérêt pour la phase d'analyse et l'application d'un algorithme d'extraction. Dans cette étape de nettoyage des données, nous avons premièrement corrigé des erreurs dans les données et inféré des valeurs manquantes. Puis, nous avons analysé certaines valeurs d'attributs numériques (typiquement, des durées) et étudié la possibilité de les regrouper en intervalles pour permettre leur prise en compte lors de l'application de l'algorithme d'extraction. Nous avons donc créé dans nos cas d'application des intervalles pour les attributs contenant des dates ou des durées, des intervalles étant plus aptes à la généralisation que si l'on analyse une valeur précise pour chaque intervention.

Avant l'application de l'algorithme *Apriori*, nous allons transformer les feuilles Excel nettoyées et mettre en forme les données d'entrée de l'algorithme pour l'extraction des règles (sous forme tabulaire). Nous construisons donc un contexte formel  $B^* = (T, I, R)$  pour chaque cas d'application, dans lequel les lignes représentent les transactions  $T$  (chacune des interventions), les colonnes représentent les *items*  $I$  (représentées par des types de concepts et leurs marqueurs définis dans le support ( $S$ )) et la relation binaire entre  $T$  et  $I$  correspond à  $R$ . Nous présentons dans la suite à titre d'exemple le contexte formel des cas # 1 et # 2.

Nous illustrons sur le Tableau IV.2 le contexte formel créé pour le **cas d'application # 1**, en tenant compte de 12 attributs d'intérêt de la base de données (nous n'allons pas prendre en compte ici le numéro de l'ordre de travail, celui-ci n'ayant pas d'intérêt pour l'analyse, toutes les interventions ayant un numéro différent) : le type d'équipement, la caractéristique de l'équipement, la zone d'emplacement, la zone de défaut, la cause de l'intervention, l'action effectuée, le type d'intervention, le poste affecté, le type de pointage, la date de l'intervention, la durée, le nom des techniciens. Ce contexte formel  $B^*$  représente la base de données d'entrée du processus de *data mining*, les transactions correspondent à l'ensemble d'interventions de maintenance ( $T = \{Intervention_1, Intervention_2, Intervention_3, \dots\}$ ), l'ensemble d'*items* correspond aux attributs (et leurs valeurs) sélectionnés pour l'analyse et contenus dans le support ( $I = \{pont1, pont2, \dots, 1crochet, 2crochet, \dots\}$ ) et  $R \subseteq T * I$  est une relation binaire entre  $T$  et  $I$  (Tableau IV.2).

Pour l'exécution d'un premier cycle d'extraction de connaissances à partir de la base de données du **cas d'application # 2**, nous avons sélectionné 15 attributs parmi les 19 présents à l'aide du responsable de maintenance. Nous illustrons sur le Figure IV.3 le contexte formel créé pour ce deuxième cas.

Comme nous l'avons souligné précédemment, l'attribut « numéro ordre de travail » est différent pour chaque intervention, donc il ne va pas être retenu dans les règles extraites, qui recherchent des points communs entre les interventions. A l'opposé, les attributs « Equipe » et « Technologie » ont toujours la même valeur : ils interviendraient donc dans toutes les règles et ont été éliminés. Une

autre colonne dans le fichier du cas # 2 est appelée « Libellé ». La valeur de cet attribut est bien identifiée quand il s'agit d'une intervention de type préventif, mais lorsqu'il s'agit d'une intervention corrective, les libellés sont très divers (il s'agit en effet d'un champ libre). On s'aperçoit que ce champ est en pratique utilisé pour entrer des symptômes (e.g. « problème de fuite sur X ») mais aussi parfois des actions (e.g. « remplacement roue chariot »). Nous avons aussi éliminé cette colonne lors de l'extraction de règles décrite dans ce mémoire, tout en commençant un travail de nettoyage de ces informations pour nos futurs travaux. En effet, pour générer des connaissances permettant d'améliorer le diagnostic ou l'efficacité des actions correctives, il serait intéressant de séparer ce champ en trois : symptôme (donnant la raison de l'intervention de la maintenance), cause (origine de la défaillance) et action menée (solution implantée). Cela demande bien sûr un travail important de retour sur ces expériences à l'aide du responsable de la maintenance (cf. les perspectives évoquées).

Tableau IV.2. Illustration du contexte formel (Cas # 1)

R	Equipement				Caractéristique		Zone d'emplacement			Zone de défaut					Cause				Actions				...
	Pont 1	Pont 2	Pont 3	...	1 crochet	2 crochets	C01	C02	...	Alimentation générale	Cabine	Commande	Direction	...	Absence signalétique	Arrêt d'urgence enclenche	Défaut angulaire	...	Amélioration fonctionnement	Essais	Modification	...	...
Intervention 1	x	o	o	o	o	x	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	...
Intervention 2	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	...
Intervention 3	o	o	o	x	o	x	o	o	x	o	o	o	o	o	o	o	o	o	o	o	o	o	...
Intervention 4	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	...
Intervention 5	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	o	x	o	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Tableau IV.3. Illustration du contexte formel (Cas # 2)

R	Date de saisi				Fréquence				Etat				Production				Nature				Priorité				...
	Juillet-Aout 2010	Sept-Déc 2010	Janv-Avril 2011	...	Annuelle	Bimestrielle	Hebdomadaire	...	En cours	En préparation	Soldé	Validé	Site1	Site2	Site3	...	AME	CAST	COR	...	Qualité	Sans urgence	Sécurité	...	...
Intervention 1	x	o	o	o	o	o	o	x	o	o	x	o	o	o	o	x	o	o	o	x	o	x	o	...	...
Intervention 2	x	o	o	o	o	o	o	x	o	o	x	o	o	o	o	x	o	o	o	x	o	x	o	...	...
Intervention 3	x	o	o	o	o	o	o	x	o	o	x	o	o	o	o	x	o	o	o	x	o	x	o	...	...
Intervention 4	x	o	o	o	o	o	o	x	o	o	x	o	o	o	o	x	o	o	o	x	o	x	o	...	...
Intervention 5	x	o	o	o	o	o	o	x	o	o	x	o	o	o	o	x	o	o	o	x	o	x	o	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...



#### IV.3.4.2 Découverte des itemsets fréquents et des règles d'association

A partir des contextes formels que nous avons construits dans la partie précédente pour chaque cas d'application, nous allons extraire des règles d'association liant les différentes valeurs d'attributs choisis qui décrivent les interventions. Rappelons que les règles d'association ne dénotent qu'une présence simultanée de plusieurs valeurs d'attributs, et ne traduisent pas une « implication ».

Nous avons choisi le logiciel SPMF<sup>9</sup> (*Sequential Pattern Mining Framework*), un logiciel open source de *data mining* écrit en langage Java par Philippe Fournier-Viger de l'Université de Moncton (Canada), dédié à l'extraction des règles d'association à partir d'un contexte formel. Pour l'extraction des différents types de règles, SPMF fournit une extension de l'algorithme *Apriori* (Agrawal & Srikant, 1994) permettant d'extraire les règles d'association à partir d'un *minsup*, d'un *minconf* et d'un *contexte formel*, qui est inclus dans le package de ce logiciel (Figure IV.10).

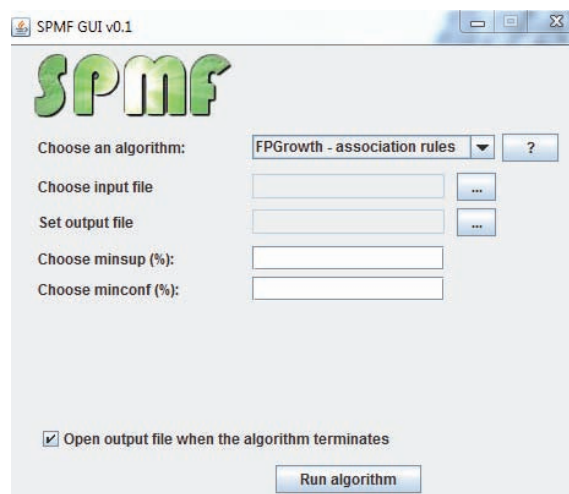


Figure IV.10. Interface du logiciel SPMF

Pour contrôler les règles extraites, nous avons donc deux indicateurs : *minsup* et *minconf*. Ces deux indicateurs classiques, ou mesures objectives, sont pris en compte lors des étapes principales de l'algorithme : le *minsup* dans la découverte des *itemsets* fréquents (Algorithme I.1) et le *minconf* dans la découverte des règles d'association (Algorithme I.2).

Rappelons ici que le support fait référence à la fréquence d'occurrence de l'hypothèse et de la conclusion par rapport au nombre total d'expériences analysées ; i.e. la règle  $A \rightarrow B$  ( $sup = 20\%$ ) peut être interprétée comme «  $A$  et  $B$  sont présents ensemble dans 20 % des expériences ». La confiance donne le pourcentage de cas dans lequel l'hypothèse, quand elle est présente, a donné lieu à la présence de la conclusion ; i.e. la règle  $A \rightarrow B$  ( $conf = 20\%$ ) veut dire que « lorsque  $A$  est présent,  $B$  est présent dans 20% des expériences contenant  $A$  ».

<sup>9</sup> <http://www.philippe-fournier-viger.com/spmf/>



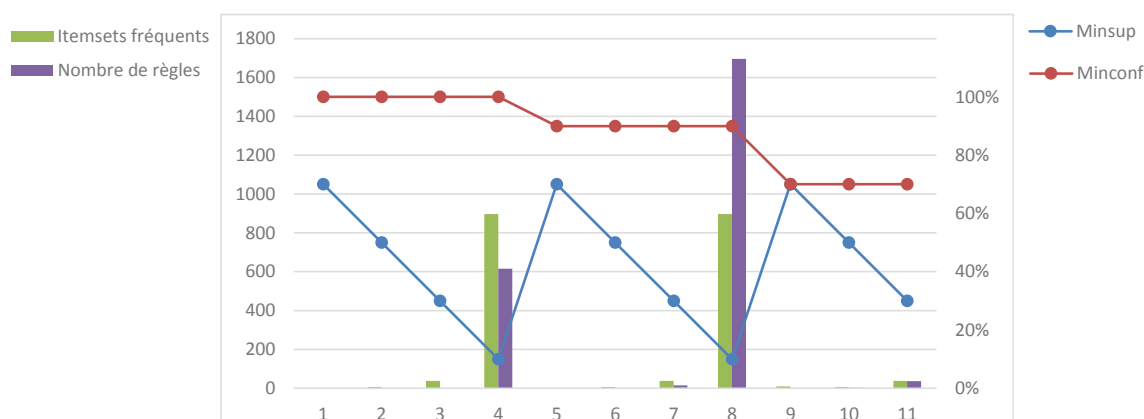
Nous présentons ci-dessous quelques résultats extraits de nos trois cas d'études réels lorsque le *minsup* et le *minconf* varient. Nous pouvons remarquer l'évolution du nombre d'*itemsets* fréquents et de règles avec la variation de ces paramètres dans le Tableau IV.4, le Tableau IV.5 et le Tableau IV.6 (respectivement), que nous allons analyser plus en détail dans la suite.

Les résultats issus du **cas d'application # 1** sont présentés dans le Tableau IV.4. Nous avons effectué 11 essais combinant différentes valeurs de *minsup* et *minconf*, en considérant 1067 interventions décrites par 12 attributs. Nous présentons dans les tableaux l'essai réalisé, le *minsup* et *minconf* défini, le nombre d'*itemsets* fréquents détectés et le nombre de règles extraites. Par exemple, dans l'essai # 7 (Tableau IV.4), un *minsup* = 30% et un *minconf* = 90% ont conduit à l'extraction de 38 *itemsets* fréquents et de 15 règles.

**Tableau IV.4. Résultats obtenus - 1067 interventions et 12 attributs (Cas # 1)**

FPGrowth - associations rules											
Essai	1	2	3	4	5	6	7	8	9	10	11
Minsup	70%	50%	30%	10%	70%	50%	30%	10%	70%	50%	30%
Minconf	100%	100%	100%	100%	90%	90%	90%	90%	70%	70%	70%
Itemsets fréquents	1	7	38	897	1	7	38	897	10	7	38
Règles extraites	0	0	1	615	0	1	15	1696	0	3	37

Nous illustrons ces résultats issus du cas # 1 dans le graphique de la Figure IV.11 afin de mieux comprendre l'influence des mesures *minsup* et *minconf* sur les *itemsets* fréquents et les règles. L'échelle pour les valeurs du *minsup* et du *minconf* est donnée sur l'axe des ordonnées de droite de la Figure IV.11, tandis que l'échelle pour le nombre d'*itemsets* et le nombre de règles apparaît sur l'axe des ordonnées de gauche. L'axe des abscisses représente les essais effectués en faisant varier le *minsup* et le *minconf* présentés dans le Tableau IV.4.



**Figure IV.11. Evolution des résultats obtenus lorsque le *minsup* et le *minconf* varient (Cas # 1)**

En général, en ayant un *minconf* constant et en diminuant le *minsup* (courbe rouge et courbe bleue respectivement sur le graphique), nous voyons une augmentation considérable du nombre d'*itemsets* et de règles (Figure IV.11). On peut constater dans les essais # 4 et # 8 qu'un faible *minsup* (ici *minsup* = 10%) conduit à de nombreux *itemsets* fréquents (représentés par les barres en vert). Au contraire, il y a une réduction drastique du nombre d'*itemsets* fréquents quand le *minsup* augmente (Dorn et al., 2008), i.e. la valeur du *minsup* contrôle directement le nombre d'*itemsets* fréquents.

Un *minsup* et un *minconf* importants conduisent à des règles moins nombreuses mais plus robustes (aucune règle dans certains cas, comme dans les essais # 1, # 5 et # 9), toutefois ces règles (dont le nombre est représenté par les barres en violet sur le graphique) ont une fréquence d'occurrence importante et une forte probabilité conditionnelle ; l'hypothèse et la conclusion sont presque toujours associées. En effet, la confiance affecte directement le nombre de règles générées. Un autre facteur important, qui affecte aussi l'évolution et l'évaluation des résultats obtenus, est la taille de l'ensemble de données, qui devrait être prise en compte lors de la définition des niveaux optimaux des indicateurs prédéfinis.

Les résultats issus du **cas d'application # 2** sont présentés dans le Tableau IV.5. Nous avons effectué 9 essais combinant différentes valeurs de *minsup* et *minconf*, en considérant les 5955 interventions décrites par 15 attributs. Nous pouvons remarquer quelques différences par rapport au cas # 1. Par exemple, le cas # 2 génère de nombreux *itemsets* fréquents et de nombreuses règles d'association lorsque *minsup* = 30% et *minconf* = 90% (essai # 7). En fait, dans ce deuxième cas, le volume de données est plus important ; de plus, de nombreux « *items* » contenus dans la base de données initiale ne varient pratiquement jamais. Pourtant, ces résultats obtenus sont sûrement peu intéressants car déjà bien connus par l'expert du domaine. Nous avons donc réalisé une nouvelle analyse, qui sera présentée ultérieurement, sans tenir compte de ces attributs très répétitifs afin de fournir à l'utilisateur de nouvelles connaissances sur l'ensemble d'interventions passées (Tableau IV.10).

**Tableau IV.5. Résultats obtenus - 5955 interventions et 15 attributs (Cas # 2)**

FPGrowth - associations rules									
Essai	1	2	3	4	5	6	7	8	9
<b>Minsup</b>	90%	50%	30%	20%	70%	50%	30%	50%	30%
<b>Minconf</b>	100%	100%	100%	100%	90%	90%	90%	80%	80%
Itemsets fréquents	2	15	101	430	3	15	101	15	101
Règles extraites	0	2	138	885	2	11	437	15	530

Nous illustrons également les résultats issus du cas # 2 dans le graphique de la Figure IV.12. De manière générale, nous pouvons remarquer dans ce cas que l'évolution du nombre d'*itemsets* et de règles d'association concorde avec les résultats issus du cas # 1.

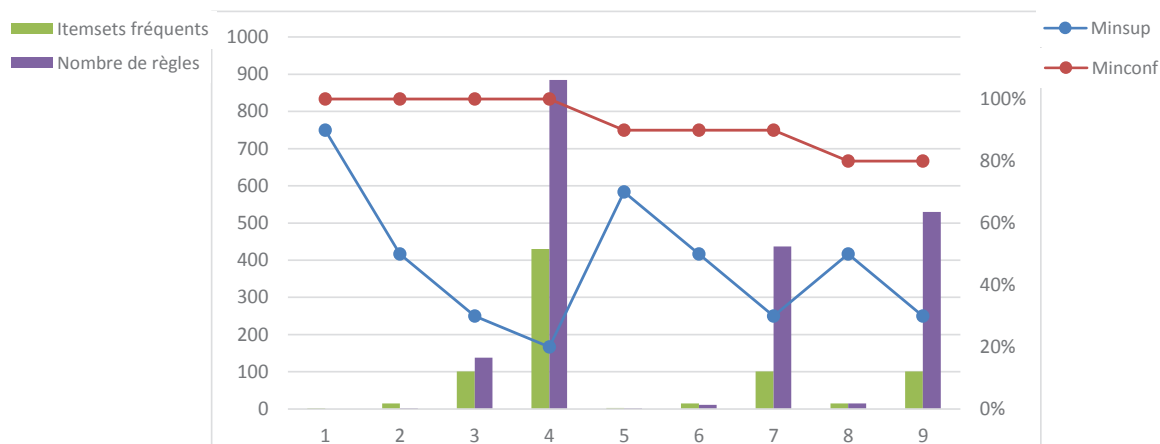


Figure IV.12. Evolution des résultats obtenus lorsque le *minsup* et le *minconf* varient (Cas # 2)

Les résultats obtenus à partir du **cas d'application # 3** sont proposés dans le Tableau IV.6. Nous avons réalisé ici 9 essais combinant différentes valeurs de *minsup* et *minconf*, en considérant les 537 interventions et 16 attributs correspondants. Nous avons testé également notre démarche sur cette base de données ayant moins d'interventions que les deux cas d'applications précédents.

Tableau IV.6. Résultats obtenus - 537 interventions et 16 attributs (Cas # 3)

FPGrowth - associations rules									
Essai	1	2	3	4	5	6	7	8	9
Minsup	90%	50%	30%	70%	50%	30%	70%	50%	30%
Minconf	100%	100%	100%	90%	90%	90%	80%	80%	80%
Itemsets fréquents	3	15	83	9	15	83	9	15	83
Règles extraites	1	23	107	8	20	150	13	31	179

#### IV.3.4.3 Evaluation objective et sémantique des règles extraites

##### Interprétation et visualisation des résultats

Cette partie porte notamment sur l'évaluation objective et sémantique des règles extraites, i.e. sur l'interprétation et la visualisation des résultats obtenus en suivant pas-à-pas la méthodologie proposée dans le Chapitre III (section III.4.2). Nous soulignons ici l'importance de la collaboration avec l'expert du domaine afin de mieux comprendre l'intérêt des règles extraites, permettant ainsi de tirer des conclusions dans le domaine.

Nous présentons dans cette partie les premiers résultats partiels de notre étude, qui pourront être améliorés si des précisions sont apportées aux bases de données considérées, et si certaines interprétations peuvent notamment être affinées (ou contredites) par les acteurs de la maintenance.

- **Analyse et interprétation des règles de base**

Pour illustrer les premières étapes de la méthodologie concernant l'analyse et l'interprétation des « règles de base » (étapes (a) à (f) - section III.4.2.1), nous allons combiner l'évaluation objective et l'évaluation sémantique en utilisant notre méthodologie pour chaque cas d'application.

### Cas d'application # 1

Dans un premier temps, nous avons choisi empiriquement les valeurs *minsup* = 30% et *minconf* = 90% pour le premier cas d'application (Tableau IV.4 - essai # 7), conduisant à l'extraction de 38 *itemsets* fréquents et de 15 règles d'association qui sont présentées dans le Tableau IV.7.

**Tableau IV.7. Règles d'association extraites (*minsup* = 30%, *minconf* = 90%) - Cas # 1**

Règle #	Règle	Sup (%)	Conf (%)
R1	Action=Essais → Type pointage=CO	31,7	98,8
R2	Zone de défaut=Translation → Type pointage=CO	40,4	96,2
R3	Type d'intervention=Correctif Urgent (CU) → Poste=Vide	40,2	94
R4	Zone d'emplacement=C02 → Type pointage=CO	42,7	93,4
R5	Zone d'emplacement=C02 → Caractéristique de l'équipement=2 crochets	45,6	99,7
R6	Type d'intervention=Assistance Technique (EC) → Type pointage=CO	45,8	99,5
R7	Caractéristique de l'équipement=2 crochets → Type pointage=CO	60,5	91,2
R8	Type d'intervention=Correctif Urgent (CU), Durée=[0-3] → Poste=Vide	33,7	96,2
R9	Type d'intervention=Correctif Urgent (CU), Type pointage=CO → Poste=Vide	35,1	93,7
R10	Zone d'emplacement=C02, Durée=[0-3] → Caractéristique de l'équipement=2 crochets	30,3	99,6
R11	Zone d'emplacement=C02, Caractéristique de l'équipement=2 crochets → Type pointage=CO	42,7	93,6
R12	Zone d'emplacement=C02, Type pointage=CO → Caractéristique de l'équipement=2 crochets	42,7	100
R13	Zone d'emplacement=C02 → Caractéristique de l'équipement=2 crochets, Type pointage=CO	42,7	93,4
R14	Type d'intervention=Assistance Technique (EC), Caractéristique de l'équipement=2 crochets → Type pointage=CO	33,5	99,4
R15	Durée=[0-3], Caractéristique de l'équipement=2 crochets → Type pointage=CO	40,1	90,1

Dans ce cas, les sept premières règles établies par l'algorithme sont considérées comme « règles de base », car elles mettent en relation deux *items* (Tableau IV.7). Comme énoncé dans le Chapitre III, l'interprétation du problème associé et l'utilité possible des règles obtenues peuvent être mieux évaluées en examinant les questions auxquelles les attributs des *items* fréquents pourraient répondre. Dans ce cas particulier, les 15 règles trouvées incluent 8 attributs donnant des réponses aux questions présentées dans le Tableau IV.8.

Tableau IV.8. Questions associées aux attributs présents - Cas # 1

Attribut :	Question :
○ Caractéristique de l'équipement	○ Combien de crochets <sup>10</sup> comporte l'équipement ?
○ Zone d'emplacement	○ Où se localise l'équipement ?
○ Zone de défaut	○ Où se localise le problème ?
○ Action	○ Quelle action a été menée ?
○ Type d'intervention	○ Dans quelles conditions ?
○ Poste	○ Quel poste est affecté ?
○ Type pointage	○ Quelle classe d'intervention a été réalisée ?
○ Durée	○ Combien de temps a duré l'intervention ?

Pour chacune des sept premières règles, nous devons exprimer le problème abordé en associant les questions correspondantes, interpréter ce problème en tenant compte des mesures d'intérêt et tirer des conclusions de la règle. Nous proposons ci-dessous l'interprétation de R1, R2, R3 et R5.

**R1 : Action=Essais → Type pointage=CO (correctif)  $sup = 31,7\%$   $conf = 98,8\%$**

*Question associée* : Y a-t-il un lien entre l'« action menée » et la « classe d'intervention réalisée » ?

*Interprétation* : Ces deux items (« action=essais » et « type pointage=CO ») sont souvent présents dans la base de données ( $sup = 31,7\%$ ). En effet, les interventions concernant des « essais » sont presque toujours considérées dans l'entreprise comme des opérations de type correctif « CO », ce qui correspond à un type de pointage défini dans la base de données initiale ( $conf = 98,8\%$ ). Seules 1,2% des interventions concernant des essais ont un autre type de pointage.

*Utilité* : Cette règle permet de savoir que les essais sont souvent liés à la classe d'intervention CO (corrective), mais pas toujours. Une explication possible de cette relation réside dans le fait que ces essais concernent des actions d'analyse effectuées pour déterminer les causes des interventions correctives dans l'entreprise. Cependant, cette classe d'intervention peut concerner d'autres types d'actions, étant donné que la règle inverse (Type pointage=CO → Action=Essais) est absente en tenant compte des mesures d'intérêt choisies. En effet, dans au moins 10% des interventions ayant un type de pointage « correctif », une action différente des essais a été réalisée (la confiance de la règle absente étant inférieure au seuil de confiance fixé ici, i.e. 90%).

**R2 : Zone de défaut=Translation → Type pointage=CO (correctif)  $sup = 40,4\%$   $conf = 96,2\%$**

*Question associée* : Y a-t-il un lien entre la « localisation du problème » et la « classe d'intervention réalisée » ?

<sup>10</sup> Accessoire d'appareil de levage qui est constitué d'une pièce recourbée qui sert à accrocher une charge à lever ou à suspendre un accessoire ou un appareil de levage à un support.

*Interprétation* : Cette règle permet de relier la zone défectueuse de l'équipement objet de l'intervention et la classe d'intervention à réaliser. La zone de translation et le pointage correctif sont souvent liés dans la base de données ( $sup = 40,4\%$ ), i.e. 40,4% des interventions analysées concernent ces deux *items*. 96,2% des interventions ayant comme zone défectueuse la zone de translation concernent une classe d'intervention de type correctif, i.e. seules, 3,8% des interventions dans la zone de translation concernent une autre classe d'intervention.

*Utilité* : Cette règle indique qu'un pourcentage important des interventions réalisées sur les ponts roulant sont correctives, et ont comme cause un problème dans la zone de translation. Il serait donc intéressant de mettre en place une stratégie de maintenance préventive plus adaptée dédiée à cette zone des ponts roulant afin de diminuer le nombre d'interventions correctives réalisées. En outre, l'absence de la règle inverse (*Type pointage*=CO  $\rightarrow$  *Zone de défaut*=Translation) a été constatée (elle n'a pas une confiance supérieure ou égale au *minconf* prédéfini (le seuil de *minconf* fixé ici pour générer une règle étant de 90%)). Cela nous permet de déduire que les interventions de type correctif ne concernent pas que la zone de translation, i.e. d'autres zones du pont ont subi aussi des interventions de type correctif, dans au moins 10% des cas (la confiance de la règle inverse étant inférieure à 90%).

**R3** : *Type d'intervention*=Correctif Urgent (CU)  $\rightarrow$  *Poste*=Vide  $sup = 40,2\%$   $conf = 94\%$

*Question associée* : Y a-t-il un lien entre les « conditions » de l'intervention et le « poste affecté » ?

*Interprétation* : Le type d'intervention correctif urgent affecte souvent un poste qui n'a pas été saisi lors de l'intervention : dans de nombreux cas, l'attribut poste est vide. En effet, 40,2% des interventions analysées correspondent à cette situation. Plus particulièrement, 94% des interventions concernant ce type d'intervention « CU » n'indiquent pas le poste affecté ; 6 % de ce type d'interventions affectent un poste qui a été saisi.

*Utilité* : Il serait intéressant de porter une attention particulière aux interventions correctives urgentes, particulièrement critiques, et de bien saisir les informations qui caractérisent ces interventions. Un point à aborder avec le responsable de maintenance est notamment de savoir pourquoi ces interventions correctives urgentes ne sont souvent pas affectées à un poste particulier. Il peut s'agir d'un problème de saisie de la part des opérateurs qui ne prennent pas le temps de saisir correctement certaines informations du fait de l'urgence de la situation. Une autre explication possible serait que pour ces interventions urgentes, une zone spécifique soit réservée mais ne soit pas indiquée dans les rapports de maintenance.

**R5** : *Zone d'emplacement*=C02  $\rightarrow$  *Caractéristique de l'équipement*=2 crochets  $sup = 45,6\%$   $conf = 99,7\%$

*Question associée* : Y a-t-il un lien entre la « localisation de l'équipement » et le « nombre de crochets comportant l'équipement » ?

*Interprétation* : Ces deux *items* (« *zone d'emplacement*=C02 » et « *caractéristique de l'équipement*=2 crochets ») sont souvent liés dans la base de données analysée ( $sup = 45,6\%$ ). 99,7% des interventions dans la zone C02 concernent des ponts avec deux crochets.

*Utilité* : Cette règle permet de connaître la caractéristique principale des ponts localisés dans cette zone ; i.e. 2 crochets. Ce type de règle peut permettre de mieux préparer une intervention de maintenance en fonction de l'endroit dans lequel elle survient, en tenant compte des caractéristiques principales des équipements qui s'y trouvent. Dans ce cas, les équipements contenant 2 crochets de cette zone correspondent aux types de pont « Pont 2 » et « Pont 3 » dans la base de données.

## Cas d'application # 2

Nous présentons ci-dessous l'interprétation de quelques règles extraites du cas d'application # 2.

Dans un premier temps, nous avons considéré les 15 attributs sélectionnés et nous avons choisi empiriquement les valeurs de *minsup* = 50% et *minconf* = 90% (Tableau IV.5 – essai # 6), menant à l'extraction de 15 *itemsets* fréquents et de 11 règles d'association qui sont présentées dans le Tableau IV.9. Parmi ces règles, les sept premières peuvent être considérées comme règles de base. Dans un deuxième temps, nous avons modifié le contexte formel issu de la base initiale, car certains attributs (colonnes) contenant presque toujours des valeurs identiques ne sont pas nécessaires. Leur retrait permettra de faire émerger d'autres règles.

**Tableau IV.9. Règles d'association (*minsup* = 50%, *minconf* = 90%) - Cas #2 (Test 1)**

Règle #	Règle	Sup (%)	Conf (%)
R1	% de réalisation=100% → Nature=PRV	52,8	95,1
R2	% de réalisation=100% → Etat=Soldé	55,5	100
R3	Section analytique=22/339/333 → Priorité=Sans urgence	51,2	91,2
R4	Section analytique=22/339/333 → Etat=Soldé	52,3	93,3
R5	Nature=PRV → Etat=Soldé	63,5	94,2
R6	Etat=Soldé → Priorité=Sans urgence	83,8	92,7
R7	Priorité=Sans urgence → Etat=Soldé	83,8	90,1
R8	% de réalisation=100%, Etat=Soldé → Nature=PRV	52,8	95,1
R9	% de réalisation=100%, Nature=PRV → Etat=Soldé	52,8	100
R10	% de réalisation=100% → Etat=Soldé, Nature=PRV	52,8	95,1
R11	Nature=PRV, Priorité=Sans urgence → Etat=Soldé	55,8	94,2

Nous remarquons d'abord que dans ces règles, on trouve présents les attributs « % de réalisation », « nature », « état », « section analytique » et « priorité » concernant les *items* fréquents détectés dans ce premier test en considérant tous les attributs (15 attributs disponibles).

Nous présentons ci-dessous l'interprétation des règles R2, R6 et R7 issues de la base de données initiale afin d'illustrer cette partie, en tenant compte du problème abordé, des mesures d'intérêt et de l'utilité suggérée.



**R2 : % de réalisation=100% → Etat=Soldé**

*sup* = 55,5% *conf* = 100%

*Question associée* : Y a-t-il un lien entre le « pourcentage d'exécution de l'ordre de travail » et l' « état de l'ordre de travail » ?

*Interprétation* : 55,5% des interventions ont été réalisées à 100% et sont dans l'état « soldé ». Toutes les interventions avec un pourcentage de réalisation de 100% concernent un état soldé (*conf* = 100%).

*Utilité* : Cette règle est « évidente » pour l'expert du domaine, toutefois, le fait que la règle inverse est absente peut amener des suggestions. La règle inverse (i.e. *Etat=Soldé* → % de réalisation=100%) n'a pas été générée, ce qui veut dire que toutes les interventions ayant un état « soldé » n'ont pas un pourcentage de réalisation de 100% ; i.e. il y a des interventions soldées avec d'autres pourcentages de réalisation (ce qui peut être vérifié dans la base) dans au moins 10% des cas (le seuil de confiance fixé ici pour générer une règle étant de 90%). Cela dénote une anomalie dans la base de données : les interventions devraient être terminées avant d'être considérées comme soldées. Une suggestion peut être de détecter les incohérences éventuelles entre état et pourcentage d'achèvement d'un ordre pour améliorer la saisie.

**R6 : Etat=Soldé → Priorité=Sans urgence**

*sup* = 83,8% *conf* = 92,7%

**R7 : Priorité=Sans urgence → Etat=Soldé**

*sup* = 83,8% *conf* = 90,1%

*Question associée* : Y a-t-il un lien entre l' « état de l'ordre de travail » et la « priorité de l'intervention » ?

*Interprétation* : Selon la règle R6, 83,8% des interventions analysées sont sans urgence et sont soldées. 92,7% des interventions en état soldé ont comme priorité « sans urgence », c'est-à-dire que 7,3% des interventions en état soldé ont eu un autre type de priorité (par exemple « urgent »).

*Utilité* : Cette règle nous indique que la plupart des interventions décrites dans la base de données ne sont pas urgentes et qu'elles ont été soldées, ce qui permet dans une certaine mesure de valider la politique de maintenance préventive de l'entreprise. L'interprétation du support pour la règle R7 est la même que pour R6, le support étant une mesure symétrique. Nous pouvons remarquer que la confiance de la règle R7 est inférieure à la confiance de R6, ce qui est lié au fait que la probabilité qu'une intervention soit sans urgence est un peu supérieure à la probabilité qu'elle soit en état soldé, du fait par exemple que des interventions « sans urgence » sont encore en préparation ou en cours (i.e. comme elles ne sont pas terminées, elles ne sont pas soldées) Il est aussi possible qu'elles soient mal renseignées dans la base. Toutefois, ces interprétations de la confiance dans ce cas n'apportent, à notre avis, aucune connaissance nouvelle à l'utilisateur.

Afin de trouver d'autres associations plus intéressantes à partir des ordres de travail (OT) du cas d'application # 2, nous n'avons ensuite plus pris en compte les attributs très répétitifs (« état », « priorité », « famille », « % de réalisation » et « section analytique »). Les dix attributs restant sont donc : la date de saisie de l'ordre de travail, la fréquence, la production, la nature, le type d'équipement, la matricule, le modèle, le temps prévu, le temps réalisé, la date de fin saisie. Nous présentons dans le Tableau IV.10 les nouveaux résultats obtenus lorsque le *minsup* et *minconf*



varient, en considérant les mêmes 5955 interventions seulement décrites par 10 attributs. Nous pouvons comparer les résultats obtenus par rapport au Tableau IV.5. Notamment, le nombre d'*itemsets* détectés et de règles extraites est inférieur en tenant compte des mêmes mesures d'intérêt, car nous avons enlevé des attributs très répétitifs. Par exemple, l'essai # 3 considérant un  $minsup = 30\%$  et un  $minconf = 100\%$  dans les deux tests a conduit ici à l'extraction de 5 règles d'association contre 138 obtenues lors du premier test (Tableau IV.5).

**Tableau IV.10. Nouveaux résultats obtenus - 5955 interventions et 10 attributs (Cas # 2)**

FPGrowth - associations rules									
Essai	1	2	3	4	5	6	7	8	9
Minsup	90%	50%	30%	20%	70%	50%	20%	50%	30%
Minconf	100%	100%	100%	100%	90%	90%	90%	80%	80%
Itemsets fréquents	0	1	13	131	0	1	30	1	13
Règles extraites	0	0	5	44	0	0	16	0	13

Pour illustrer ces nouveaux résultats issus du cas # 2, nous avons choisi empiriquement un  $minsup = 20\%$  et un  $minconf = 90\%$  (Tableau IV.10 – essai # 7), conduisant à l'extraction de 30 *itemsets* fréquents et de 16 règles d'association présentées dans le Tableau IV.11.

Avec ces valeurs de  $minsup$  et de  $minconf$  définis, les six premières règles établies sont considérées comme règles de base. Nous présentons ci-dessous l'interprétation des règles R1 et R3.

**R1 : Fréquence=Semestrielle  $\rightarrow$  Nature=PRV**

$sup = 20,8\%$   $conf = 99,9\%$

*Question associée* : Y a-t-il un lien entre la « fréquence de l'intervention » et le « type d'intervention » ?

*Interprétation* : 20,8% des 5955 interventions analysées sont réalisées tous les 6 mois et sont de type préventif (PRV). Presque toutes les interventions ayant une fréquence semestrielle correspondent à une intervention préventive ( $conf = 99,9\%$ ).

*Utilité* : Cette règle est évidente et n'apporte aucune connaissance nouvelle à l'expert du domaine, toutefois elle confirme qu'il n'y a pas d'anomalies dans la base de données par rapport à ces deux *items*. Dans ce cas, la règle inverse est absente parce qu'une intervention préventive peut en effet avoir une fréquence différente selon l'équipement considéré, par exemple hebdomadaire, annuelle, etc.

**R3 : Temps prévu=[1:00, 2:00]  $\rightarrow$  Nature=PRV**

$sup = 30,1\%$   $conf = 95,1\%$

*Question associée* : Y a-t-il un lien entre le « temps prévu » et le « type d'intervention » ?

*Interprétation* : 30,1% des interventions analysées ont une durée prévue entre 1 et 2 heures et sont préventives. 95,1% des interventions avec cette durée prévue sont de nature préventive, c'est-à-dire que seules 4,9% des interventions de cette durée concernent un autre type d'intervention (correctif ou correctif suite préventif par exemple).

*Utilité* : Cette règle nous permet de savoir que les interventions préventives avec des temps d'intervention prévus entre 1 et 2 heures sont fréquentes (30,1%) dans la base données considérée. Cette règle permet de mieux planifier les interventions à réaliser en fonction de la durée prévue. La règle inverse (*Nature=PRV* → *Temps prévu*=[1:00, 2:00]) est absente, ce qui confirme que les interventions préventives ont aussi d'autres durées prévues.

**Tableau IV.11. Règles d'association (*minsup* = 20%, *minconf* = 90%) - Cas #2 (Test 2)**

Règle #	Règle	Sup (%)	Conf (%)
R1	<i>Fréquence</i> =Semestrielle → <i>Nature</i> =PRV	20,8	99,9
R2	<i>Production</i> =CF332 → <i>Type équipement</i> =X100	23,3	96,7
R3	<i>Temps prévu</i> =[1:00, 2:00] → <i>Nature</i> =PRV	30,1	95,1
R4	<i>Modèle</i> =CABINES → <i>Type équipement</i> =PEINTURE	35,2	100
R5	<i>Modèle</i> =CABINES → <i>Production</i> =Site2	35,2	100
R6	<i>Type équipement</i> =PEINTURE → <i>Production</i> =Site2	38,9	98,1
R7	<i>Modèle</i> =CABINES, <i>Nature</i> =PRV → <i>Type équipement</i> =PEINTURE	27,7	100
R8	<i>Type équipement</i> =PEINTURE, <i>Production</i> =Site2 → <i>Modèle</i> =CABINES	35,2	90,5
R9	<i>Modèle</i> =CABINES, <i>Production</i> =Site2 → <i>Type équipement</i> =PEINTURE	35,2	100
R10	<i>Modèle</i> =CABINES, <i>Type équipement</i> =PEINTURE → <i>Production</i> =Site2	35,2	100
R11	<i>Modèle</i> =CABINES → <i>Production</i> =Site2, <i>Type équipement</i> =PEINTURE	35,2	100
R12	<i>Modèle</i> =CABINES, <i>Nature</i> =PRV → <i>Production</i> =Site2	27,7	100
R13	<i>Type équipement</i> =PEINTURE, <i>Nature</i> =PRV → <i>Production</i> =Site2	30,9	99,5
R14	<i>Modèle</i> =CABINES, <i>Production</i> =Site2, <i>Nature</i> =PRV → <i>Type équipement</i> =PEINTURE	27,7	100
R15	<i>Modèle</i> =CABINES, <i>Type équipement</i> =PEINTURE, <i>Nature</i> =PRV → <i>Production</i> =Site2	27,7	100
R16	<i>Modèle</i> =CABINES, <i>Nature</i> =PRV → <i>Production</i> =Site2, <i>Type équipement</i> =PEINTURE	27,7	100

### Cas d'application # 3

Comme dans les deux cas précédents, nous n'avons pas pris en compte l'attribut « numéro de l'ordre de travail » dans le cas d'application # 3, puisque chaque intervention possède un numéro unique d'OT ; cette information n'est donc pas pertinente dans l'analyse.

Les premiers résultats décrits ci-dessous pourront être améliorés avec une participation plus active des experts du domaine lors de la phase de prétraitement des données. Par exemple, une meilleure saisie des champs renseignés de manière « libre » pourrait en particulier permettre de générer des connaissances permettant d'améliorer le diagnostic ou le choix des actions correctives. Dans la base de données disponible, il serait intéressant de revoir les champs tels que « description » concernant le libellé de l'OT, « résumé de la défaillance », et « description de la cause racine », pour bien distinguer les symptômes (raison de l'intervention de la maintenance), les causes (origine de la défaillance) et les actions menées (solution implantée).

Nous présentons avec plus de détails l'interprétation que nous pouvons faire de quelques règles générées lors d'un premier test, sans prendre en considération les attributs renseignés de manière « libre ».

Dans un premier temps, nous allons prendre en compte 16 attributs pour l'analyse. Nous avons remarqué lors du premier test que certains *items* dans la base ne varient pratiquement jamais, ce qui conduit à la découverte de connaissances certainement évidentes pour les experts du domaine. Nous avons donc mené un deuxième test sans tenir compte des attributs contenant presque toujours des valeurs identiques, afin d'obtenir de nouveaux résultats.

Dans le premier test, mené avec un *minsup* = 40% et un *minconf* = 100%, nous avons obtenu 27 *itemsets* fréquents et 25 règles d'association. Nous ne présentons dans le Tableau IV.12 que les règles de base issues du test (une liste complète est présentée dans l'Annexe 2).

**Tableau IV.12. Règles d'association (*minsup* = 40%, *minconf* = 100%) - Cas # 3 (Test 1)**

Règle #	Règle	Sup (%)	Conf (%)
R1	<i>Nature de l'objet</i> =Piping & fittings → <i>Groupe de planification</i> =PU1	41,7	100
R2	<i>Nature de l'objet</i> =Piping & fittings → <i>Site</i> =FRXX	41,7	100
R3	<i>Durée de la défaillance</i> =0-2h → <i>Site</i> =FRXX	42	100
R4	<i>Type d'ordre</i> =PM11 → <i>Site</i> =FRXX	70,5	100
R5	<i>Unité</i> =A01 → <i>Groupe de planification</i> =PU1	87,7	100
R6	<i>Unité</i> =A01 → <i>Site</i> =FRXX	87,7	100
R7	<i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	98,5	100
...	...	...	...

Nous remarquons ici l'extraction de règles très « robustes », toutes ayant une confiance de 100%. Parmi les 16 attributs analysés, on trouve les 6 attributs suivants présents dans l'ensemble de règles extraites : « nature de l'objet », « groupe de planification », « site », « durée de la défaillance », « type d'ordre » et « unité ». Pour faciliter l'interprétation de ces règles, nous avons de nouveau associé une question à chacun de ces attributs (Tableau IV.13).

**Tableau IV.13. Questions associées aux attributs présents - Cas # 3 (Test 1)**

Attribut :	Question :
○ <i>Nature de l'objet</i>	○ Quel type d'objet ?
○ <i>Groupe de planification</i>	○ Planifié dans quelle catégorie ?
○ <i>Site</i>	○ Dans quel site ?
○ <i>Durée de la défaillance</i>	○ Combien de temps ?
○ <i>Type d'ordre</i>	○ Quel type d'ordre ?
○ <i>Unité</i>	○ Dans quelle unité ?

Pour chacune de ces sept premières règles, nous devons exprimer le problème abordé, l'interprétation des mesures d'intérêt dans ce contexte et l'analyse de l'utilité possible de la règle. Nous présentons ci-dessous à titre d'exemple l'interprétation de la règle R4.

**R4** : Type d'ordre=PM01 → Site=FRXX

*sup* = 70,5% *conf* = 100%

*Question associée* : Y a-t-il un lien entre le « type d'ordre » et le « site » ?

*Interprétation* : 70,5% des ordres de travail sont sur le site FRXX et de type PM01, correspondant aux pannes d'urgence 2 & 3 (ne bloquant pas la production). Les interventions de ce type sont toujours réalisées sur le site FRXX (*conf* = 100%).

*Utilité* : Cette règle n'apporte aucune nouvelle connaissance à l'expert du domaine puisqu'un seul site « *site*=FRXX » est considéré dans la base de données analysée. Toutefois, elle nous permet de connaître la probabilité des interventions de type PM11, qui sont très fréquentes dans l'entreprise (70,5%). Afin de faire émerger des associations plus intéressantes pour l'utilisateur, nous avons réalisé une nouvelle analyse, cette fois sans considérer ces *items* très répétitifs.

Afin de trouver d'autres règles à partir des OT du fichier du cas d'application # 3, nous n'avons ensuite plus pris en compte les attributs « type d'ordre », « site », « unité » et « groupe de planification ». Nous présentons dans le Tableau IV.14 les nouveaux résultats obtenus avec les 12 attributs restants, en considérant le même nombre de transactions.

**Tableau IV.14. Nouveaux résultats obtenus - 537 interventions et 12 attributs (Cas # 3)**

FPGrowth - associations rules									
Essai	1	2	3	4	5	6	7	8	9
Minsup	50%	10%	8%	15%	10%	20%	10%	20%	10%
Minconf	100%	100%	100%	70%	70%	50%	50%	30%	30%
Itemsets fréquents	0	107	161	41	107	18	107	18	107
Règles extraites	0	1	3	5	26	4	60	4	166

En comparant les résultats du Tableau IV.14 avec ceux du Tableau IV.6, nous pouvons identifier des différences concernant le nombre d'*itemsets* fréquents et de règles d'association obtenues. Dans ce test, nous avons dû prendre en compte des valeurs de *minsup* plus faibles afin d'obtenir un nombre significatif de règles. Néanmoins, nous avons défini lors d'un essai, les mêmes seuils *minsup* et *minconf* dans les deux tests (avec 16 puis 12 attributs). Nous remarquons que lorsque le *minsup* = 50% et le *minconf* = 100%, nous n'avons pas obtenu d'*itemsets* fréquents, et donc aucune règle d'association (Tableau IV.14 – essai # 1) alors que dans le test précédent (test 1), ces mêmes seuils ont abouti à 15 *itemsets* fréquents et 23 règles d'association (Tableau IV.6 – essai # 2).

Nous avons choisi ici un *minsup* = 10% et *minconf* = 70% pour illustrer les nouveaux résultats, menant à l'extraction de 107 *itemsets* fréquents et de 26 règles d'association (Tableau IV.14 – essai # 5). Dans le Tableau IV.15, nous ne présentons que les règles de base obtenues (voir l'Annexe 3 pour une liste complète).

Tableau IV.15. Règles d'association (*minsup* = 10%, *minconf* = 70%) - Cas # 3 (Test 2)

Règle #	Règle	Sup (%)	Conf (%)
R1	Groupe d'équipement=DIV → Criticité de l'équipement=#N/A	11,5	100
R2	Description de l'objet=#N/A → Date de création=sept-oct-nov-dec	12,1	98,4
R3	Groupe d'équipement=100 → Criticité de l'équipement=Medium	13,4	83,7
R4	Effet de la défaillance=Vide → Nature de l'objet=Vide	13,5	76,8
R5	Nature de l'objet=Vide → Effet de la défaillance=Vide	13,5	70,1
R6	Effet de la défaillance=Vide → Type de cause=Vide	14,8	84,2
R7	Type de cause=Vide → Nature de l'objet=Vide	16	71,6
R8	Nature de l'objet=Vide → Type de cause=Vide	16	82,6
R9	Groupe d'équipement=500 → Criticité de l'équipement=High	16,3	73,9
R10	Type de cause=Wear and tear → Nature de l'objet=Piping & fittings	19,9	77,5
R11	Effet de la défaillance=Leak → Nature de l'objet=Piping & fittings	23,2	75,7
...	...	...	...

Les questions auxquelles les attributs des *items* fréquents dans les règles pourraient répondre sont exposées dans le Tableau IV.16.

Tableau IV.16. Questions associées aux attributs présents- Cas # 3 (Test 2)

Attribut :	Question :
○ Description de l'objet	○ quel équipement ?
○ Date de création	○ quels mois de l'année ?
○ Groupe d'équipement	○ quel groupe d'équipement ?
○ Criticité de l'équipement	○ quel niveau de criticité ?
○ Effet de la défaillance	○ quel effet ?
○ Nature de l'objet	○ quel type d'objet ?
○ Type de cause	○ quelle cause ?

Nous présentons ci-dessous l'interprétation que nous pouvons faire de quelques règles générées (R3, R4, R5 et R10) :

**R3** : Groupe d'équipement=100 → Criticité de l'équipement=Intermédiaire (*medium*)

*sup* = 13,4% *conf* = 83,7%

*Question associée* : Y a-t-il un lien entre le « groupe d'équipement » et le « niveau de criticité » ?

*Interprétation* : 13,4% des interventions concernent le groupe d'équipement 100 avec une criticité de niveau intermédiaire. 83,7% des interventions sur les équipements du groupe 100 ont une criticité intermédiaire, donc 16,3% des interventions concernant un équipement de ce groupe correspond aussi à d'autres niveaux de criticité.

*Utilité* : Cette règle permet de savoir que la plupart des équipements du groupe 100 ont une criticité intermédiaire. De telles règles peuvent permettre une meilleure préparation des activités de maintenance. La règle inverse (*Criticité de l'équipement*=Intermédiaire → *Groupe d'équipement*=100) n'a pas été générée avec les seuils prédéfinis, ce qui nous permet de déduire que ce niveau de criticité affecte d'autres groupes d'équipements dans l'entreprise dans au moins 30% des cas (le seuil de confiance fixé ici pour générer une règle étant de 70%).

**R4** : *Effet de la défaillance*=Vide → *Nature de l'objet*=Vide *sup* = 13,5% *conf* = 76,8%

**R5** : *Nature de l'objet*=Vide → *Effet de la défaillance*=Vide *sup* = 13,5% *conf* = 70,1%

*Question associée* : Y a-t-il un lien entre l' « effet » et le « type d'objet » ?

*Interprétation* : 13,5% des interventions analysées n'ont pas de valeurs saisies pour les attributs « effet de la défaillance » et « nature de l'objet ». D'une part, 76,8% des interventions qui ne donnent pas l'effet de la défaillance ne décrivent pas non plus la nature de l'objet (règle R4). D'autre part, 70,1% des interventions qui ne donnent pas la nature de l'objet ne donnent pas non plus l'effet de la défaillance (règle R5).

*Utilité* : Ces deux règles nous montrent des anomalies contenues dans la base de données analysée et suggèrent de lancer des actions pour mieux saisir les expériences. En effet, ces attributs essentiels dans la description d'une intervention devraient être toujours remplis afin d'avoir une meilleure traçabilité des interventions.

**R10** : *Type de cause*=Usure (*Wear and tear*) → *Nature de l'objet*=Tuyauteries et raccords (*Piping & fittings*) *sup* = 19,9% *conf* = 77,5%

*Question associée* : Y a-t-il un lien entre la « cause » et le « type d'objet » ?

*Interprétation* : 19,9% des interventions analysées ont eu comme type de cause l'usure des tuyauteries et raccords. 77,5% des interventions renfermant ce type de cause concernent les tuyauteries et raccords, donc 22,5% des interventions causées par l'usure ont affecté un autre type d'objet.

*Utilité* : Cette règle indique que l'usure est l'origine principale des interventions réalisées sur les tuyauteries et raccords. Il pourrait être intéressant dans ce cas de revoir les plans de maintenance préventive sur ces objets afin de mieux analyser le phénomène d'usure et d'éviter que celle-ci ne conduise à une défaillance. La règle inverse (*Nature de l'objet*=Tuyauteries et raccords → *Type de cause*=Usure) n'a pas été générée, ce qui veut dire que les interventions ayant comme objet les tuyauteries et raccords ont aussi des causes de défaillances différentes de l'usure (ce qui peut être vérifié dans la base) dans au moins 30% des interventions (le seuil de confiance fixé ici pour générer une règle étant de 70%).

- **Analyse des règles plus spécifiques**

Dans cette étape, nous considérons les règles plus spécifiques, qui correspondent à des variantes des règles de base que nous venons de présenter et d'analyser, obtenues par trois opérations logiques : *extension*, *permutation* et *jonction* (étape (g) - section III.4.2.2).

**Extensions de la partie hypothèse ou conclusion des règles :** comme nous l'avons souligné dans le Chapitre III, de nouveaux *items* peuvent être ajoutés à la partie hypothèse ou conclusion de la règle, ce qui signifie que la règle devient plus spécifique et que les mesures d'intérêt peuvent être influencées. Nous allons présenter des exemples issus des trois cas d'application.

Dans le **cas d'application # 1**, nous avons trouvé des extensions des sept premières règles de base (Tableau IV.7). Par exemple, la règle R8 est une extension de la partie hypothèse de la règle R3 (i.e. un nouvel *item* a été ajouté dans la partie hypothèse). Dans ce cas, la règle R8 a un support plus faible que R3 et une confiance un peu plus forte, indiquant ainsi l'influence sur les indicateurs support et confiance de l'*item* « *Durée*=[0-3] » dans la nouvelle règle. Nous présentons ci-dessous une analyse de la nouvelle règle obtenue (R8).

**R3 :** *Type d'intervention*=Correctif Urgent (CU) → *Poste*=Vide sup=40,2% conf=94%

**R8 :** *Type d'intervention*=Correctif Urgent (CU), *Durée*=[0-3] → *Poste*=Vide sup=33,7% conf=96,2%

*Question associée :* Y a-t-il un lien entre les « conditions », la « durée » et le « poste affecté » ?

*Interprétation :* 33,7% des interventions analysées concernent des interventions de type correctif urgent, ayant eu une durée inférieure à 3 heures et sans saisie du poste affecté. Les interventions de caractère urgent avec cette durée affectent presque toujours (96,2%) un poste qui n'a pas été saisi lors de l'intervention. Seules, 3,8% des interventions de ce type ont un poste affecté saisi.

*Utilité :* La comparaison des mesures d'intérêt de ces deux règles indique que l'absence de la mention du poste affecté est encore plus probable pour les interventions correctives urgentes et courtes (i.e. d'une durée inférieure à 3 heures), ce qui pourrait confirmer une des hypothèses formulées lors de l'interprétation de la règle R3 : comme les opérateurs n'ont pas beaucoup de temps pour réaliser l'intervention du fait de l'urgence de la situation, ils ne prennent pas le temps de saisir le poste. Par rapport à la règle R3, nous pouvons déduire de cette nouvelle règle que moins d'interventions correctives urgentes ont des durées d'intervention supérieures à 3 heures, car la fréquence d'occurrence conjointe des *items* « *Type d'intervention*=Correctif Urgent (CU) » et « *Durée supérieure à 3 heures* » est inférieure à celle d'avoir ce type d'intervention et l'*item* « *Durée inférieure à 3 heures* ». Cette information est obtenue en comparant les supports des deux règles analysées.



En outre, nous remarquons dans le Tableau IV.7 que R9 est une extension de R3 (partie hypothèse) ; R10 une extension de R5 ; R11 une extension de R4 et de R7 ; R12 de R5 ; R13 de R4 et de R5 ; R14 de R6 et de R7 ; R15 de R7, etc. Nous allons mieux représenter ces relations dans l'étape suivante de la méthodologie.

Concernant les résultats issus du **cas d'application #2** (test 2), nous allons analyser les variantes des six règles de base trouvées quand le *minsup* = 20% et le *minconf* = 90% (Tableau IV.11). La règle R7 (*Modèle*=CABINES, *Nature*=PRV → *Type équipement*=PEINTURE) est par exemple une extension de la partie hypothèse de R4 (*Modèle*=CABINES → *Type équipement*=PEINTURE). Ici, R7 a une fréquence d'apparition plus faible, conservant toutefois la même confiance que R4. La règle R9 (*Modèle*=CABINES, *Production*=Site2 → *Type équipement*=PEINTURE) est aussi une extension de la partie hypothèse de R4, conservant le même support et la même confiance que la règle élémentaire. Cette nouvelle règle (R9), plus informative car elle donne le site dans lequel sont réalisées les interventions, devrait remplacer la règle R4.

D'autres extensions ont été trouvées dans ce deuxième cas d'application telles que R10 (extension de R5 et de R6) ; R11 (extension de R4 et de R5) ; R12 de R5 ; R13 de R6 ; R14 (extension de R4, de R7 et de R9) ; R15 (extension de R5, de R6, de R10, de R12 et de R13) ; R16 (extension de R4, de R5, de R7, de R11 et de R12).

Dans le **cas d'application #3**, nous prenons aussi en compte le deuxième test réalisé ne concernant que 12 attributs. Ce test a généré 26 règles d'association parmi lesquelles les 11 premières sont des règles de base (Tableau IV.15). Parmi les règles extraites, R12 (*Effet de la défaillance*=Vide, *Type de cause*=Vide → *Nature de l'objet*=Vide) est une extension des parties hypothèses de R4 (*Effet de la défaillance*=Vide → *Nature de l'objet*=Vide) et de R7 (*Type de cause*=Vide → *Nature de l'objet*=Vide) ; R13 est une extension de R6 et de R8 ; R14 de R5 ; R15 de R4 et de R6 ; R16 de R6 ; R18 de R7 ; R19 de R8.

**Permutations des items entre la partie hypothèse et conclusion des règles** : des permutations des items entre la partie hypothèse et la partie conclusion des règles peuvent survenir dans les résultats. Dans ces cas, le support des règles est identique mais le niveau de confiance peut être différent.

Pour le **cas d'application #1**, par exemple la règle R12 est une permutation de la règle R11 (Tableau IV.7). Dans ce cas, l'item « *type de pointage*=CO » (correctif) qui est dans la partie conclusion de R11 et l'item « *caractéristique de l'équipement*=2 crochets » qui est dans l'hypothèse ont été permutés dans la règle R12. R13 est aussi une permutation de R11 et de R12. Ces trois règles, présentées ci-dessous, présentent un même support étant donné qu'il s'agit d'une mesure symétrique (voir section I.5.3.3). Leurs confiances varient un peu : il est donc possible d'en tirer quelques conclusions. Il n'y a que la règle R12 qui ait une confiance=100%, ce qui signifie que les items présents dans l'hypothèse permettent de conclure avec certitude sur la présence des items de la partie la conclusion. Nous interprétons dans la suite la règle R12 et analysons quelques différences avec ses permutations R11 et R13.



**R11** : Zone d'emplacement=C02, Caractéristique de l'équipement=2 crochets → Type pointage=CO  
*sup* = 42,7% *conf* = 93,6%

**R13** : Zone d'emplacement=C02 → Caractéristique de l'équipement=2 crochets, Type pointage=CO  
*sup* = 42,7% *conf* = 93,4%

**R12** : Zone d'emplacement=C02, Type pointage=CO → Caractéristique de l'équipement=2 crochets  
*sup* = 42,7% *conf* = 100%

*Question associée* : Y a-t-il un lien entre la « localisation de l'équipement », la « classe d'intervention réalisée » et le « nombre de crochets comportant l'équipement » ?

*Interprétation* : 42,7% des interventions analysées concernent des interventions correctives sur les ponts à 2 crochets de la zone C02 (cette interprétation est la même pour les règles R11 et R13). Concernant la confiance, cette règle R12 nous indique que les interventions correctives sur les équipements de la zone C02 sont toujours réalisées sur les équipements ayant 2 crochets (*conf* = 100%).

*Utilité* : L'hypothèse de cette règle permet de conclure sur la caractéristique du pont roulant (2 crochets), i.e. elle permet d'identifier quels équipements d'une zone spécifique il faut préparer pour les interventions correctives. Toutefois, la règle R11 apporte d'autres connaissances intéressantes ; par exemple, R11 indique implicitement que les équipements de la zone C02 ayant 2 crochets peuvent aussi subir des interventions de classes autres que correctives, dans 6,4% des cas. Dans ces cas, l'interprétation de la confiance de chaque règle peut apporter des nouvelles connaissances à l'utilisateur. Nous ne pouvons pas donner ici des lignes directrices pour choisir quelle règle devrait être retenue par rapport aux autres (chaque règle fournit en effet une information différente) ; nous donnons néanmoins ici quelques pistes pour mieux interpréter les règles extraites notamment selon les besoins de l'utilisateur et le type d'information qu'il recherche dans les données.

Nous remarquons que pour le **cas d'application # 2** (test 2), il existe aussi quelques permutations. Par exemple, les règles R9, R10 et R11 sont des permutations de R8, faisant passer l'*item* « *modèle*=CABINES » de la conclusion de R8 aux parties hypothèses des nouvelles règles (Tableau IV.11). Ces quatre règles ont le même support (*sup* = 35,2%), toutefois, seule la règle R8 a une confiance inférieure à 100%. Dans ce cas, les règles R9, R10 et R11 n'apportent aucune nouvelle connaissance à l'utilisateur (par rapport à la confiance) l'ordre des attributs n'ayant aucune incidence sur l'indicateur de la confiance. Nous pouvons a priori considérer que les règles permutées, ayant la même confiance, peuvent ne pas être prises en compte : elles sont redondantes.

L'interprétation de la règle R8 peut donner toutefois des informations intéressantes.

**R8** : Type équipement=PEINTURE, Production=Site2 → Modèle=CABINES *sup* = 35,2% *conf* = 90,5%

*Question associée* : Y a-t-il un lien entre le « type d'équipement », le « site » et le « modèle » ?

*Interprétation* : 35,2% des interventions analysées sont réalisées sur les cabines du site2 et correspondent à l'activité de peinture (cette interprétation du support de la règle est la même pour R9, R10 et R11). 90,5% des interventions concernant la peinture sur le site2 concernent les

cabines, 9,5% des interventions concernant la peinture sur ce site étant liées à un autre « modèle ».

*Utilité* : Cette règle nous confirme que les cabines sont l'objet de la plupart des interventions concernant la peinture réalisées sur le site<sup>2</sup>. La règle permet de formaliser qu'il y a aussi d'autres modèles liés à l'activité de peinture qui ont lieu sur ce site.

Les autres permutations du cas d'application # 2 correspondent aux règles R14, R15 et R16, toutes ayant même support et même confiance (Tableau IV.11). Par rapport à la confiance, ces règles n'apportent aucune nouvelle connaissance à l'utilisateur, en effet les *items* mis en relation sont indépendants et leur ordre n'affecte pas les mesures d'intérêt prises en compte ici : il s'agit de règles redondantes. L'unique information issue de ces règles concerne leurs fréquences d'apparition ; en effet, 27,7% des interventions lient cabines et peinture sur le site<sup>2</sup> de l'entreprise, ces interventions étant de type préventif.

Concernant le **cas d'application # 3**, nous pouvons remarquer certaines permutations entre règles. Par exemple, la règle R5 (*Nature de l'objet*=Vide → *Effet de la défaillance*=Vide) est une permutation de R4 (*Effet de la défaillance*=Vide → *Nature de l'objet*=Vide), la conclusion de R4 devenant l'hypothèse de R5 et vice-versa (Tableau IV.15). L'analyse de cette permutation permet de déduire quelques informations. Par rapport au support, et comme nous l'avons déjà dit, ces deux règles indiquent que 13,5% des interventions n'ont pas de valeurs saisies pour les attributs « effet de la défaillance » et « nature de l'objet ». Toutefois, nous pouvons déduire de la confiance de la règle R4 que 23,2% des interventions n'ayant pas d'attribut « effet de la défaillance » saisi concernent une autre valeur pour l'attribut « nature de l'objet », i.e. le type d'objet est donné.

En outre, R7 est une permutation de R8. Toutefois, R8 a une confiance plus élevée. Les règles R12, R13, R14 et R15 (voir l'Annexe 3) ont toutes un support de 13% concernant les mêmes *items* : 13% des interventions de maintenance dans cette base de données n'ont pas de valeurs saisies pour les attributs « effet de la défaillance », « type de cause », « nature de l'objet ». Ces règles ont des valeurs de confiance différentes, montrant l'influence des *items* dans la partie hypothèse sur ceux qui sont dans la partie conclusion. Elles ne sont donc pas redondantes. D'autres permutations détectées concernent R16 avec R17, et R18 avec R19.

**Jonctions des parties hypothèse ou des parties conclusion des règles** : certaines règles combinent la partie hypothèse ou la partie conclusion de deux autres règles. Nous présentons ci-dessous quelques exemples.

Pour le **cas d'application # 1** par exemple, la règle R11 est une jonction des règles R4 (*Zone d'emplacement*=C02 → *Type pointage*=CO) et R7 (*Caractéristique de l'équipement*=2 crochets → *Type pointage*=CO), ce qui revient à dire, comme souligné précédemment, que la règle R11 est une extension de R4 mais aussi de R7. Dans ce cas, les parties hypothèses des règles R4 et R7 se combinent dans la nouvelle règle (Tableau IV.7). La règle R11 ayant le même support de la règle R4, nous pouvons déduire que ce sont les probabilités de l'*item* « zone d'emplacement=C02 » et de l'*item*

« *type de pointage=CO* » qui déterminent le support des deux règles. D'autres jonctions identifiées dans ce cas sont : R13 est une jonction de R4 et de R5 ; et R14 de R6 et de R7.

Nous présentons à titre illustratif les jonctions identifiées afin de fournir à l'utilisateur des informations complémentaires sur l'origine de la nouvelle règle générée (i.e. les règles de base qui la conforment). L'utilité potentielle et l'interprétation des nouvelles connaissances apportées par les nouvelles règles a déjà été analysée dans la partie d'analyse des extensions.

Dans cette étape de la méthodologie proposée, l'analyse des liens entre les règles est importante car de nouvelles connaissances peuvent être créées en corrélant les interprétations des mesures d'intérêt (ici le support et la confiance) des règles proches, plus particulièrement les interprétations liées aux extensions des règles et aux permutations identifiées.

- **Représentation des liens entre les règles extraites**

Nous avons présenté initialement l'ensemble des règles extraites sous la forme de tableaux. Il est clair à ce stade qu'une structure globale synthétisant l'ensemble des règles sous forme graphique peut faciliter la compréhension, l'exploration et la perception de l'information contenue dans les règles et des relations existantes entre les règles.

Nous avons identifié dans l'étape précédente trois relations importantes entre les règles extraites par un algorithme comme *Apriori*. Il peut maintenant être utile de représenter ces liens pour fournir une aide à l'utilisateur final dans la phase de validation.

Afin d'illustrer cette étape (*étape (h) - section III.4.2.3*), nous présentons dans la suite à titre d'exemple l'ensemble de règles extraites issues du cas d'application # 1 (Tableau IV.7), en tenant compte des différents liens identifiés entre les règles. Dans la Figure IV.13, les sept règles de base (plus génériques) sont représentées au niveau supérieur, donnant un aperçu général du type de connaissance extraite. Les variantes des règles de base sont représentées à un niveau inférieur, décrivant ainsi des règles plus spécifiques (règles comprenant plus d'*items*) issues des règles de base par les opérations logiques définies précédemment.

La structure globale de l'ensemble de règles extraites est donc représentée par des graphes conceptuels dans lesquels un concept (sommet) de *TC* (hiérarchie des concepts) correspond à une des règles extraites, les mesures de support et confiance étant représentées comme une instance individuelle de chaque concept représentant une règle, et chaque relation de *TR* (hiérarchie des relations) représentant le lien entre deux règles. La représentation de l'opération « jonction » n'est pas explicitée dans la Figure IV.13 car comme nous l'avons dit, une règle qui est une extension de deux règles de base traduit sans doute la jonction de ces deux règles.

Cette représentation procure une meilleure facilité de lecture globale des règles, en fournissant une vue générale des relations existantes entre règles (mais sans expliciter le contenu des règles). Elle permet ainsi à l'utilisateur de sélectionner les « règles liées » à analyser pour choisir d'en éliminer certaines ou pas, sans avoir besoin d'explorer tout l'ensemble des règles extraites. Ainsi par exemple, si un utilisateur est intéressé à un moment donné par la règle R3 (en haut à gauche sur la Figure IV.13), il pourra analyser également les extensions de cette règle pour en tirer des conclusions intéressantes (les nouvelles règles peuvent être redondantes, mais elles peuvent aussi apporter des

informations complémentaires). L'utilisateur doit pour cela interpréter les règles par rapport à sa sémantique métier et évaluer leur utilité. Cette structure complète donc l'information fournie dans le Tableau IV.7 et permet d'organiser les connaissances dans la base de règles.

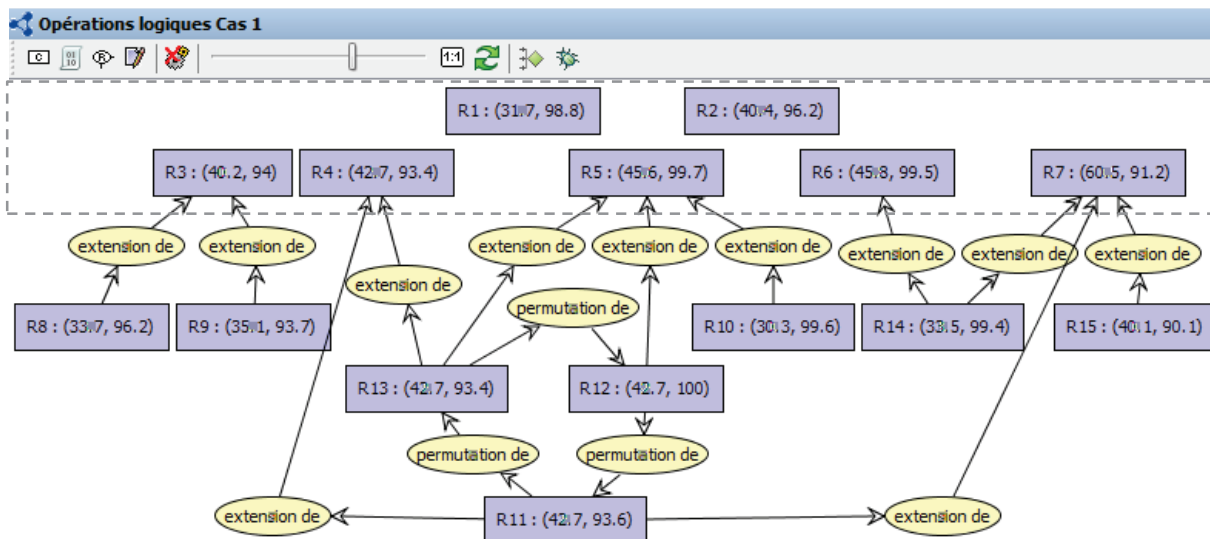
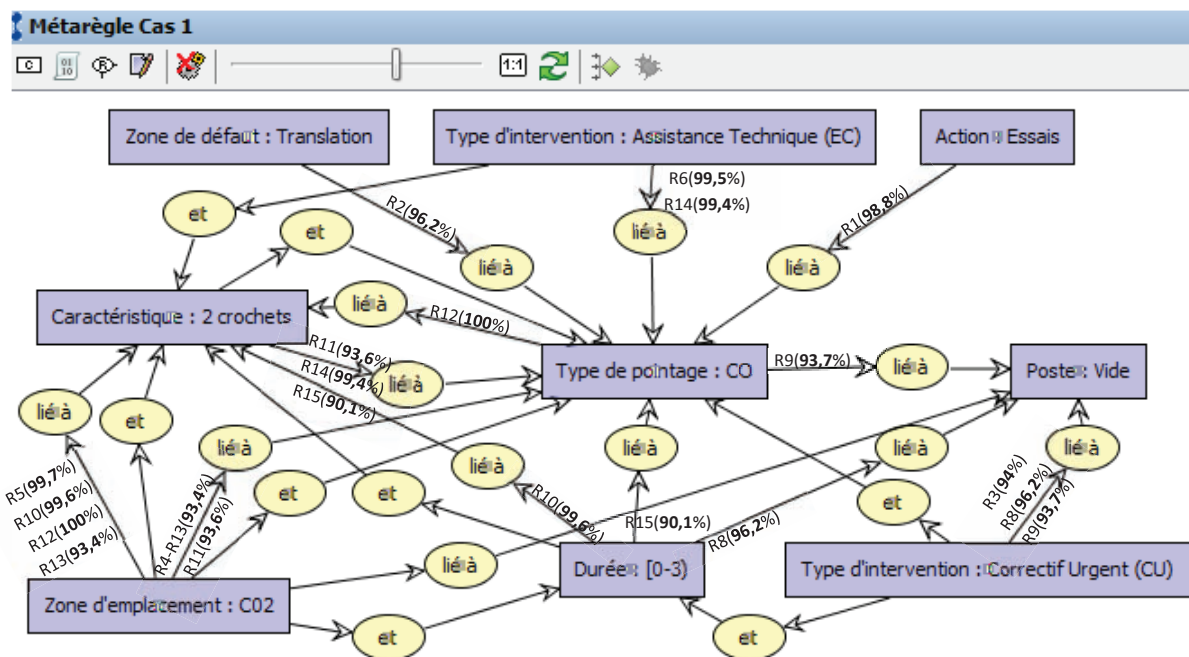


Figure IV.13. Liens entre les règles extraites (*sup, conf*) - Illustration du cas # 1

- **Représentation des liens entre les items détectés dans les règles extraites**

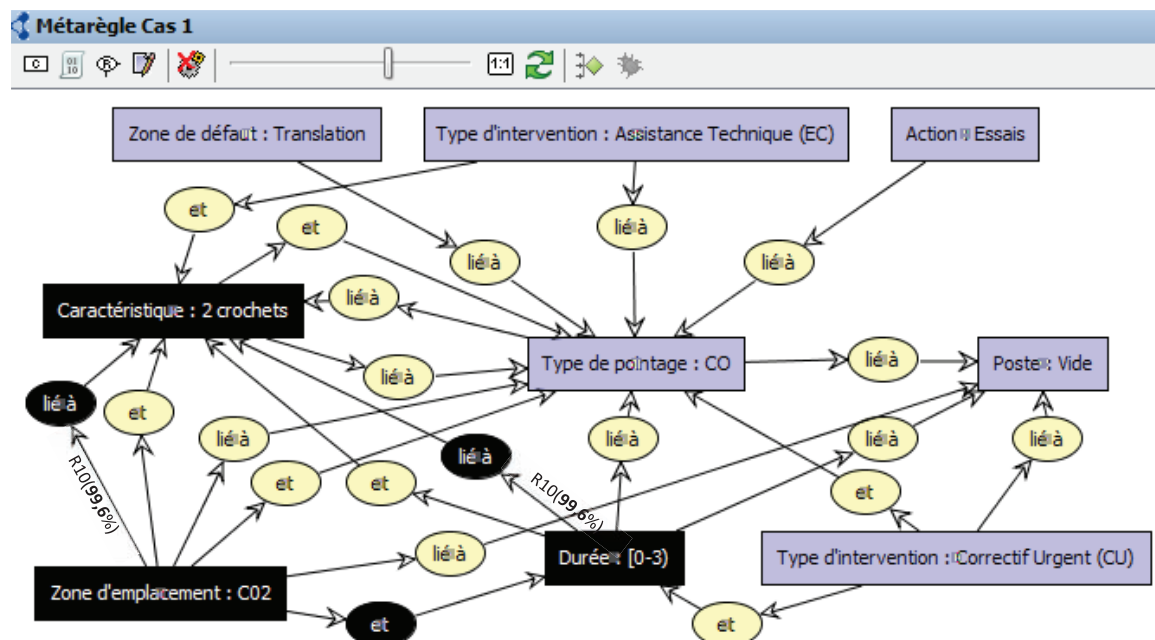
Pour illustrer l'étape (i) de la méthodologie (section III.4.2.4), les GCs sont à nouveau utilisés ici pour la formalisation d'une « métarègle » qui fournit un nouveau niveau d'abstraction plus spécifique. Cette métarègle inclut les relations existantes entre tous les *items* fréquents trouvés dans les règles extraites, facilitant la visualisation des résultats. En suivant le **cas d'application # 1**, nous présentons dans la Figure IV.14 la métarègle incluant les relations entre les *items*, de la partie hypothèse et conclusion. Dans cette figure sont contenues les 15 règles extraites du cas # 1 (Tableau IV.7).

Comme souligné précédemment, pour formaliser chacune des règles extraites, nous avons construit des GCs simples qui relient les *items* de l'hypothèse et de la conclusion par la relation « lié à », et les *items* contenus dans une même partie (hypothèse ou conclusion) par la relation « et ». Nous avons ensuite réuni dans un nouveau graphe les GCs représentant chacune des règles par l'« opération de jointure » définie dans les graphes conceptuels. Cette nouvelle représentation (Figure IV.14) complète ainsi l'information fournie dans la Figure IV.13. On peut par exemple vérifier si les règles inverses sont présentes, ainsi que la redondance entre les règles en tenant compte de leur confiance, représentée ici sur les arcs d'entrée (ou arcs amont) de chaque relation « lié à ».



**Figure IV.14. Métarègle regroupant tous les items des règles extraites et leurs liens (Cas #1)**

Pour vérifier la connaissance représentée, nous illustrons par exemple la règle R10 contenue dans la métarègle (Figure IV.15). Ce résultat est issu de la projection de R10 sur la métarègle.



**Figure IV.15. Illustration la règle R10 contenue dans la métarègle**

Cette règle R10 met en relation trois attributs concernant le contexte de l'intervention (« durée », « zone d'emplacement » et « caractéristique de l'équipement »), compte tenu du modèle générique d'une expérience présenté dans la Figure IV.8.

#### IV.3.4.4 Exploitation de la base de règles : recherche des types spécifiques de règles selon les attentes de l'utilisateur

Après avoir considéré les GCs pour la représentation des différents types de connaissances, nous allons considérer ici les mécanismes de raisonnement de ce formalisme afin d'aider le filtrage et la validation de l'ensemble de règles extraites lors de la recherche de types spécifiques de règles. A partir des visualisations de l'ensemble de règles et des relations entre les *items* contenues dans ces règles, l'utilisateur pourra donc modéliser les requêtes qui correspondent à ses attentes, puis filtrer les règles selon leur forme et/ou la présence impérative de certains *items* (concepts dans *TC*) ou de leurs spécialisations.

Nous allons présenter un exemple afin d'illustrer ce processus de filtrage. Cette évaluation subjective va toujours dépendre du type de règle que l'utilisateur souhaite trouver dans la base.

En considérant toujours les résultats issus du **cas d'application # 1** (Tableau IV.7), nous présentons dans la Tableau IV.16 une illustration d'une requête *Q* d'un utilisateur, exprimée par un graphe conceptuel décrivant le type de règle attendue dans un contexte donné. Dans cet exemple, l'utilisateur cherche parmi les règles extraites celles de la forme [Durée : \*]→(lié à)→[Type de pointage : \*]. Dans ce modèle, l'utilisateur décrit les attributs qui devraient être inclus dans la partie hypothèse et la partie conclusion, mais sans restreindre ces attributs à des valeurs spécifiques. Ainsi, l'opération de projection (traduisant la relation de subsumption) des GCs va rechercher les concepts décrits dans la requête ou leurs spécialisations selon la hiérarchie des types de concepts (*TC*) définie.

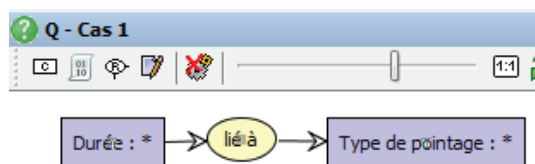


Figure IV.16. Requête (Q) de l'utilisateur

En tenant compte de la structure des règles, la requête *Q* va trier l'ensemble de règles extraites dans quatre classes de règles potentiellement intéressantes lors de l'opération de projection. Nous présentons ci-dessous quelques exemples afin d'illustrer les résultats obtenus.

#### Règles conformes

Parmi les règles extraites en tenant compte des *minsup* et *minconf* prédéfinis dans ce cas, nous n'avons pas de règles complètement conformes à la requête *Q* de l'utilisateur, i.e. aucune règle n'a

à la fois les parties hypothèse et conclusion en cohérence avec  $Q$  en tenant compte des hiérarchies  $TC$  et  $TR$  prédéfinies.

### Règles avec partie conclusion inattendue

Nous présentons dans la Figure IV.17 un exemple de règle dont la conclusion est inattendue. Nous remarquons que la partie hypothèse de la règle R8 inclut l'hypothèse de la requête, mais fournit en même temps plus d'information.

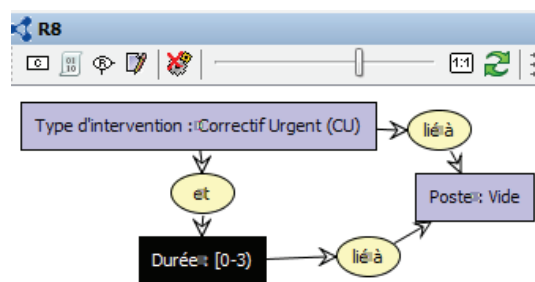


Figure IV.17. Règle R8 avec partie conclusion inattendue

### Règles avec partie hypothèse inattendue

Afin d'illustrer cette classe de règle, nous présentons ici des exemples de règles trouvées dans l'ensemble de règles extraites dont la partie hypothèse est inattendue. Dans la Figure IV.18, nous pouvons voir clairement que la partie conclusion de la règle R1 est une projection de la conclusion de  $Q$ , mais pas la partie hypothèse. La conclusion de la règle R1 est conforme à la conclusion de la requête, étant donné que « CO » est une spécialisation du concept « type de pointage » dans le support ontologique ( $S$ ).

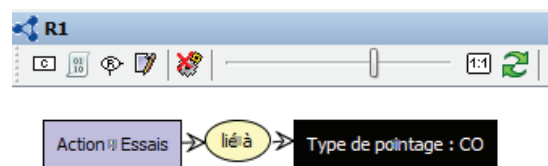


Figure IV.18. Règle R1 avec partie hypothèse inattendue

Nous pouvons remarquer aussi que la règle R15 (Figure IV.19) inclut ou est conforme avec la requête  $Q$ , toutefois la partie hypothèse de la règle fournit à l'utilisateur une information supplémentaire (ici la caractéristique du pont). Pour cela, nous considérons que la règle R15 doit être considérée dans cette catégorie.



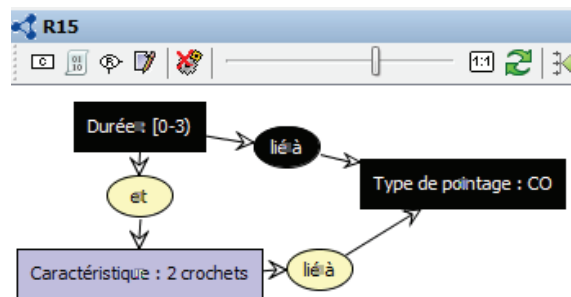


Figure IV.19. Règle R15 avec partie hypothèse inattendue

### Règles avec hypothèse et conclusion inattendues

Des règles sont inattendues par rapport à  $Q$  si à la fois leur hypothèse et leur conclusion ne sont pas des projections de  $Q$ . Nous présentons dans la Figure IV.20 un exemple d'une règle inattendue trouvé dans l'ensemble de règles considéré.

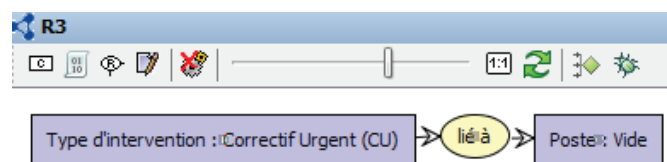


Figure IV.20. Règle R3 avec hypothèse et conclusion inattendues

## IV.4 BILAN

Dans ce dernier chapitre, nous avons présenté l'application de notre approche générale REx-ECD au domaine de la maintenance industrielle. Nous nous sommes intéressés dans cette étude aux registres (ou historiques du bien) des interventions techniques issus des outils de GMAO d'entreprises des secteurs aéronautique et pharmaceutique.

Dans un premier temps, nous avons présenté un état de l'art des systèmes ou outils d'aide à la maintenance basés sur la connaissance. Nous avons plus particulièrement décrit des systèmes basés sur des méthodes de gestion des connaissances/expériences, et d'autres prenant en compte des techniques de fouille de données, ce qui nous a permis de positionner nos travaux de recherche.

La deuxième partie de ce chapitre a été consacrée à l'analyse et au traitement de trois cas d'application réels suivant les principales étapes de notre démarche REx-ECD. Nous avons ajusté et instancié cette démarche en tenant compte des volumes très différents d'informations (interventions décrites par des attributs) mis en jeu dans chaque cas d'application.



Tout d'abord, nous avons structuré une base d'expériences formalisées basée sur un support ontologique qui vise à faciliter leur traitement et leur réutilisation. Ce support ontologique a été volontairement simplifié pour l'application de notre approche afin de rendre plus facile sa compréhension. Nous soulignons ici qu'il devrait être enrichi en fonction des caractéristiques et des restrictions de chaque entreprise, le support ontologique étant un élément déterminant pour l'expressivité des différents types de connaissances (expériences et règles) contenues dans la base de REx-ECD. Dans nos cas d'application, nous avons en effet élaboré des supports réduits à partir des informations contenues dans les bases de données considérées, et c'est en fonction de ces supports que nous avons proposé un modèle pour la représentation des expériences.

En ce qui concerne le processus d'extraction de connaissances, nous avons mis en œuvre les principales étapes nécessaires pour extraire des règles d'association à partir des informations décrivant les interventions de chaque cas. Dans ce processus, nous avons porté une attention particulière à la phase de post-traitement afin d'évaluer l'intérêt des règles extraites dans le domaine de la maintenance, sur la base de plusieurs idées originales pour :

- analyser et interpréter les connaissances extraites et l'influence des valeurs de *minsup* et *minconf* sur les résultats de chaque cas d'application,
- tirer des conclusions sur l'utilité potentielle des règles dans chaque contexte, non seulement de celles qui sont présentes, mais aussi du fait que certaines règles attendues sont absentes,
- visualiser à différents niveaux d'abstraction les liens possibles entre les différents types de règles pour faciliter la compréhension par l'utilisateur de la base de connaissances générée, et
- filtrer les règles extraites à l'aide d'un mécanisme de requête-réponse afin que l'utilisateur puisse être alimenté en connaissances adaptées à ses intérêts et/ou ses besoins du moment.

Bien que la mise en relief de l'utilité des règles demande une excellente connaissance des pratiques de maintenance à laquelle nous n'avons pas toujours eu accès, nous considérons que ces premiers résultats de l'application de notre approche dans le domaine de la maintenance sont très prometteurs. Soulignons en effet qu'une implication importante des acteurs du domaine dans le processus est un gage d'efficacité de la démarche générale suggérée, en permettant d'affiner l'impact des règles générées sur les pratiques de maintenance, point qui n'a pour l'instant été abordé que superficiellement.



## CONCLUSION ET PERSPECTIVES

---

Dans le cadre de la valorisation des expériences passées des entreprises, la problématique des méthodes de gestion des expériences abordée dans cette thèse constitue aujourd'hui un enjeu important afin d'améliorer les performances des processus industriels, de la conception des produits à leur fabrication, en passant par la maintenance des équipements de production.

Aujourd'hui, les démarches de retour d'expérience s'orientent principalement sur la collecte et la capitalisation des connaissances spécifiques attachées à une activité ou à un contexte particulier passé, dans le but de tirer des leçons des expériences passées. Dans ce contexte, nous avons abordé d'une part l'intérêt de la réutilisation de ces fragments de connaissances concrètes (les expériences) pour faciliter la résolution d'un nouveau problème ; d'autre part, nous avons étudié comment ces expériences peuvent être analysées a posteriori afin d'en extraire de nouvelles connaissances et ainsi constituer et enrichir le référentiel de connaissances des entreprises. C'est dans ce cadre que nos travaux de recherche sont positionnés.

Nous avons donc proposé d'élargir les démarches classiques de Retour d'Expérience (REx) en considérant la réalité actuelle des systèmes d'informations industriels, caractérisés par la généralisation des ERP (*Enterprise Resource Planning*). En effet, des quantités d'informations toujours plus importantes sont recueillies dans les systèmes d'information industriels, mais ne sont pas nécessairement réutilisées et/ou traitées pour analyser ce qui s'est passé et ainsi améliorer les processus. Ce constat nous a amené à nous intéresser aux techniques d'Extraction de Connaissances à partir de Données (ECD), qui permettent de générer de nouvelles connaissances à partir de l'analyse des expériences passées pour aider aux futures prises de décision.

Dans ce cadre d'amélioration du référentiel de connaissances de l'entreprise pour aider la prise de décision, cette thèse vise principalement à intégrer, de manière explicite, un processus d'Extraction de Connaissances à partir de Données (ECD) au processus de Retour d'Expérience (REx) afin de générer de nouvelles connaissances plus générales que les expériences.

La démarche conceptuelle REx-ECD proposée intègre des aspects cognitifs ainsi que statistiques afin de concevoir une mémoire d'entreprise (la base de REx-ECD) et d'allier les avantages de ces deux approches émergentes visant à capitaliser et exploiter les différents types de connaissances issues

de l'expérience dans un contexte industriel. Notre apport a principalement porté sur les étapes mises en œuvre afin de structurer une base d'expériences formalisées (*connaissances spécifiques*) concernant les informations qui caractérisent les expériences recueillies, et de générer une base de connaissances extraites sous la forme de règles (*connaissances plus génériques*) par la généralisation du résultat de nombreuses expériences. Ces bases (expériences et règles) ont en commun un modèle de la connaissance, ou support ontologique.

Une fois le cadre général posé, nous avons proposé tout d'abord une formalisation du vocabulaire du domaine et une représentation générique des expériences sous la forme : contexte, analyse et solution, à l'aide d'un formalisme de représentation de connaissances, afin de faciliter le partage et la réutilisation des expériences. De plus, nous avons mis en évidence l'intérêt des techniques relatives à la fouille de données (ou techniques de *data mining*) pour extraire de nouvelles connaissances dans un contexte de retour d'expérience. Nos travaux de recherche ont eu pour principal objectif de répondre à cette dernière problématique.

En ce qui concerne la représentation des connaissances, nous avons porté une attention particulière, tout au long de la démarche REx-ECD, aux graphes conceptuels, formalisme de représentation proposé pour supporter la méthodologie et pour modéliser les types de connaissances contenues dans la base de REx-ECD. Ce formalisme favorise ainsi une représentation unifiée de la connaissance, sa structuration sémantique, son traitement et sa compréhension par l'utilisateur, en vue d'une exploitation future.

A l'égard du processus d'ECD, nous avons considéré trois étapes principales :

- le *prétraitement* afin de préparer les données à analyser,
- le *data mining* concernant l'application des algorithmes d'extraction,
- le *post-traitement* relatif à l'évaluation des résultats obtenus.

Dans un contexte de retour d'expérience, la méthode d'extraction de règles d'association nous a paru la méthode de *data mining* la plus appropriée afin de traiter et analyser les informations décrivant les expériences recueillies. En effet, cette méthode se base sur le comptage de cooccurrence de valeurs d'attributs, c'est-à-dire des invariants identifiés dans les expériences, qu'il est ensuite possible de traduire sous la forme de règles à l'aide des indicateurs statistiques. Nous avons souligné qu'en dépit du vocabulaire utilisé dans le domaine de l'extraction des règles d'association, de telles règles ne désignent pas des implications ou liens de causalité ; elles traduisent seulement des corrélations. Les principaux avantages de la méthode d'extraction de règles d'association sont les suivants :

- le modèle des motifs extraits est relativement simple et compréhensible par un utilisateur non spécialiste en *data mining*, puisque des relations sous la forme de règles sont souvent considérées comme proches du raisonnement humain ;
- la méthode permet d'analyser de grandes bases de données, mais aussi, éventuellement, très peu de données ;
- la charge de travail des utilisateurs pendant le processus d'extraction (notamment dans la phase de *data mining*) reste limitée, sauf lors de l'évaluation et de la validation des résultats, phases dans lesquelles leur participation est décisive.

Cependant, plusieurs inconvénients et/ou limites ont été rencontrés lors de sa mise en œuvre, tels que :

- la tendance à découvrir un grand nombre de règles d'association ou le fait que toutes les règles ne sont pas pertinentes ou intéressantes pour les utilisateurs finaux ;
- l'attention nécessaire à l'interprétation des règles, relativement importante, les règles extraites pouvant s'avérer redondantes, incohérentes ou encore contradictoires ;
- la difficulté pour interpréter les règles en tenant compte des mesures d'intérêt ou des indicateurs statistiques.

Ce constat a mis en avant la nécessité d'une phase de *post-mining* afin d'améliorer la qualité de la connaissance extraite. Sur la base de cette réflexion, nous avons insisté sur l'importance d'évaluer l'intérêt des règles à l'aide d'experts du domaine avant leur réutilisation et/ou leur intégration dans le système industriel. Notre contribution en ce sens réside dans la proposition d'une démarche structurée qui prend en compte trois manières d'évaluer les règles :

- une *évaluation objective* basée sur les mesures d'intérêt associées aux règles (nous avons utilisé dans nos travaux les deux mesures classiques prises en compte dans les algorithmes d'extraction de règles : le support et la confiance),
- une *évaluation sémantique* qui considère les connaissances du domaine et la sémantique associée aux motifs extraits,
- une *évaluation subjective* qui dépend surtout du point de vue de l'expert du domaine, qui est l'utilisateur final.

Cette manière d'évaluer les règles a pris en considération des travaux déjà décrits dans la littérature. Sa nouveauté résulte cependant dans la manière d'interpréter, de visualiser et de filtrer les connaissances extraites par l'algorithme d'extraction.

En ce qui concerne les *évaluations objective et sémantique*, nous avons proposé et testé une approche visant à analyser, visualiser et comparer les règles obtenues, en suivant pas à pas une méthodologie structurée qui favorise la collaboration entre un expert de *data mining* et un expert du domaine. Cette méthode permet de garantir que les résultats extraits sont considérés comme de « nouvelles connaissances » pour aider la prise de décisions stratégiques, en particulier dans le cadre de l'amélioration continue des processus. Cette approche décrit la manière avec laquelle les règles d'association peuvent être analysées et interprétées, en tenant compte de plusieurs facteurs : l'analyse du support et de la confiance des règles, l'interprétation de la présence et/ou de l'absence de règles attendues dans les résultats, l'analyse de l'utilité de chaque règle, la visualisation des relations entre les différentes règles extraites, la proposition d'une métarègle synthétisant la connaissance extraite et regroupant les relations existantes entre les *items* des règles dans une seule structure, et l'interaction entre les experts. Concernant l'*évaluation subjective*, nous avons proposé lors de l'exploitation de la base de règles, un mécanisme de requête-réponses permettant de filtrer les règles selon le point de vue et/ou les attentes de l'expert dans une situation particulière. Dans cette étape, nous avons de nouveau profité des avantages visuels et logiques des graphes conceptuels. Ces graphes, utilisés en particulier pour la représentation des règles extraites et la visualisation de leurs relations, fournissent une aide à l'utilisateur final pour interpréter plus

facilement l'information présentée et tester différentes hypothèses en fonction de ses intérêts ou de ses intuitions.

Enfin, dans la dernière partie de ce mémoire, la démarche proposée a été appliquée à des cas réels dans le domaine de la maintenance industrielle. Ce domaine d'étude a été choisi car une grande quantité d'information décrivant les expériences passées est en général disponible du fait de la généralisation des outils de GMAO dans les entreprises, ces outils étant souvent intégrés aux ERP.

Malgré l'attention particulière que nous avons portée à ce domaine industriel, il nous paraît utile d'insister toutefois sur le fait que la démarche conceptuelle proposée est générique et peut être instanciée et adaptée dans de nombreux autres domaines et/ou départements, tels que la conception, les méthodes, la qualité ou la gestion de production, entre autres.

Ce premier bilan tiré de nos travaux de recherche, combiné aux limites identifiées lors de l'application de notre approche au domaine de la maintenance, a permis de mettre en lumière des perspectives pour les développements futurs de ces travaux. Ces perspectives pourraient passer par :

- des informations complémentaires qui seraient apportées aux bases de données considérées lors de la phase de *prétraitement* afin d'enrichir les expériences,
- le test d'autres mesures d'intérêt et algorithmes utilisés lors du *data mining*,
- un affinage des interprétations (ou leur contradiction) par les experts du domaine lors de la phase de *post-traitement*.

Ces différents points (et quelques autres) sont développés ci-dessous.

**En phase de prétraitement** (section II.3.1), le travail de nettoyage des données est décisif pour améliorer la qualité des résultats obtenus. Nous nous sommes aperçus dans les bases de données analysées dans le domaine de la maintenance que des attributs importants décrivant les interventions sont saisis de manière libre. Le manque d'homogénéité dans leur libellé a fait que nous n'avons pas pu les prendre en compte dans cette première étude. Nous pensons que ces attributs ont un fort potentiel comme sources de nouvelles connaissances, mais leur harmonisation demande un travail important afin d'homogénéiser l'information qui y est contenue. Nous avons donc commencé le fastidieux travail de nettoyage de ces informations pour nos futurs travaux afin de vérifier l'influence de ces informations sur les résultats finaux.

Dans cette même étape, nous pensons qu'il serait utile de fragmenter la base de données initiale par rapport à ses périodes d'acquisition, puis d'analyser le contenu des différentes sous-bases de manière séparée à l'aide d'un algorithme d'extraction. L'idée serait de comparer l'évolution des règles générées au cours du temps, et de tenter de corrélérer cette évolution soit à des caractéristiques intrinsèques des équipements (vieillesse par exemple), soit à des démarches d'amélioration déjà menées (plans de maintenance préventive par exemple). Il serait ainsi possible d'utiliser les règles non comme source de connaissances, mais comme indicateurs indirects de validation d'actions menées. Il s'agit donc ici d'un cas d'application particulier et simple de *clustering* (regroupement) parfois suggéré lors du prétraitement des données dans le processus d'ECD (Quan et al., 2009), mais à notre connaissance rarement utilisé avec un critère de regroupement temporel.

**En phase de *data mining*** (section II.3.2.2), et plus particulièrement lors de l'extraction des règles d'association, les algorithmes d'extraction de règles ont tendance à produire un grand nombre de règles parfois inintéressantes pour l'utilisateur. Il nous semble donc pertinent de tester différentes mesures d'intérêt. Nous avons considéré dans ce travail les deux mesures classiques proposées par l'algorithme *Apriori*, sur la base d'un critère d'éligibilité, afin d'évaluer les règles d'un point de vue statistique. Toutefois, comme nous l'avons évoqué précédemment, le support et la confiance ont des limites et/ou inconvénients qui peuvent être palliés en améliorant ces mesures ou en utilisant d'autres, par exemple, les mesures de Lift (International Business Machines, 1996) ou de conviction (Brin, Motwani, Ullman & Tsur, 1997). Il nous semblerait donc utile, dans de futurs travaux, de tester et d'expérimenter certaines de ces nouvelles mesures d'intérêt pour vérifier si elles permettent d'apporter une information complémentaire et utile lors de l'interprétation des résultats. Dans ce cadre et étant donné que nous nous sommes concentrés principalement sur l'analyse et l'évaluation des résultats obtenus par l'algorithme *Apriori*, algorithme fiable et très répandu, nous pensons également qu'il serait nécessaire de tester d'autres algorithmes d'extraction de règles d'association réputés plus performants, tels que l'algorithme FP-Growth (Han & Pei, 2000), l'algorithme « top-K » (Fournier-Viger et al., 2012), ou l'approche d'extraction IGB (Gasmi et al., 2006).

**En phase de post-traitement** (section II.3.3), une implication plus importante des experts du domaine est un gage d'efficacité de la démarche suggérée. Nous avons insisté dans cette démarche REx-ECD sur l'importance de la participation de l'expert du domaine dans le processus, surtout dans cette dernière phase, afin d'analyser et d'évaluer les nouvelles connaissances et d'adapter les techniques d'extraction de connaissances à ses besoins spécifiques. En effet, l'implication de l'expert permet d'affiner l'impact des règles générées en fonction des pratiques du domaine. Cette dernière problématique doit permettre d'établir des liens précis entre les types de règles obtenues et les améliorations potentielles des procédures et/ou des processus dans l'entreprise.

De plus, concernant toujours cette phase, nous pensons qu'il serait utile d'étudier et d'évaluer l'intérêt de mettre dans une même structure les différentes interfaces proposées lors de la visualisation des résultats, i.e. les liens entre les règles (section III.4.2.3) et liens entre tous les *items* détectés dans les règles (III.4.2.4).

**L'utilisation des ontologies déjà existantes dans l'entreprise** est une autre perspective qui nous paraît intéressante. Nous pensons que la possibilité de coupler et d'adapter notre démarche REx-ECD à des ontologies standard, par exemple dans le domaine de la maintenance, comme IMAMO (Industrial Maintenance Management Ontology) ou MIMOSA (Maintenance Information Open system), pourrait être d'un grand intérêt pour faciliter l'intégration et l'interopérabilité de la démarche REx-ECD avec les systèmes d'information industriels. Dans ce cas, une grande partie du travail à réaliser consisterait à adapter la formalisation des expériences et des règles, ainsi que la phase de prétraitement des données pour les préparer en fonction de l'ontologie existante.

**La prise en compte du caractère incertain des informations** utilisées pour extraire les règles d'association semble aussi une problématique intéressante qui pourrait apporter une information additionnelle lors de l'analyse et l'interprétation des règles extraites afin de tenir compte de la fiabilité des résultats obtenus pour la prise de décisions. Il serait donc utile de tenir compte du caractère imprécis et incomplet des bases de données lors de la phase de *data mining* afin de mesurer le degré d'incertitude associé à ces bases de données imparfaites, qui seront analysées pour

extraire les règles d'association. L'intérêt des différents formalismes permettant la représentation de l'incertitude de ces données, tels que la théorie des probabilités, la théorie des possibilités ou encore la théorie des fonctions de croyances, devrait pour cela être analysé.

Finalement, **la mise à jour ou la maintenance de la base de règles** est un dernier point décisif qui devrait être abordé et étudié afin d'estimer la périodicité du processus d'extraction de connaissances dans une entreprise et la manière avec laquelle les connaissances seront mises à jour (Tseng et al., 2007), i.e. comment les nouvelles connaissances extraites seront indexées à la base de règles existante afin de préserver, actualiser et améliorer les connaissances stratégiques de l'organisation. Pour cela, la base de données source et sa fréquence de mise à jour doivent être analysées et évaluées. Ces analyses doivent conduire à une stratégie pour extraire et indexer de nouvelles connaissances, et supprimer ou modifier les connaissances déjà existantes.

L'« entreprise basée sur la connaissance » est un thème de recherche important de plus en plus mis en avant au niveau national et international. Nos travaux s'inscrivent complètement dans ce thème, et nous espérons lui avoir apporté une contribution en participant au courant de recherche très prometteur visant à générer des connaissances à partir des informations contenues dans les ERP, encore peu exploitées.



## TABLE DE FIGURES

Figure I.1. Différentes types de connaissance.....	7
Figure I.2. Processus de création et de transfert de connaissances (Nonaka & Takeuchi, 1995).....	9
Figure I.3. Cycle classique de la gestion des connaissances selon Grundstein (Grundstein, 2004)...	11
Figure I.4. De l'événement et l'expérience à la connaissance (adaptée de (Rakoto, 2004)).....	13
Figure I.5. Structuration des processus et des composants supports du REx.....	16
Figure I.6. Intégration de données avant l'application de techniques de data mining .....	23
Figure I.7. Réduction de données avant l'application de techniques de data mining .....	24
Figure I.8. Transformation des données avant l'application de techniques de data mining.....	24
Figure II.1 : Exemple de support (adaptée de (Kamsu-Foguem & Noyes, 2013)).....	43
Figure II.2 : Exemple de GC (Kamsu-Foguem & Noyes, 2013).....	44
Figure II.3. Relation de subsomption (Kamsu-Foguem & Noyes, 2013) .....	45
Figure II.4. Illustration d'un graphe conceptuel simple et d'un graphe conceptuel emboîté .....	46
Figure II.5. Principales étapes de la démarche conceptuelle REx-ECD .....	48
Figure II.6. Illustration de la préparation des données .....	49
Figure II.7. Principales étapes pour la structuration de la base d'expériences.....	51
Figure II.8. Modèle générique pour représenter une expérience.....	52
Figure II.9. Un prétraitement avant le data mining .....	53
Figure II.10. Illustration de la transformation des données.....	54
Figure II.11. Illustration des résultats obtenus en fonction du <i>minsup</i> et du <i>minconf</i> .....	56
Figure II.12. Illustration de la génération des <i>itemsets</i> candidats et des <i>itemsets</i> fréquents .....	57
Figure II.13. Illustration de la génération des règles d'association.....	58
Figure II.14. Exploitation possible de la base de REx-ECD (adaptée de (Béler, 2008)) .....	59
Figure II.15. Une réutilisation possible de la base d'expériences formalisées.....	61
Figure III.1. Modèle conceptuel proposé pour le partage de connaissances entre l'expert du domaine et l'expert d'ECD.....	66
Figure III.2. L'intégration IHM dans notre approche conceptuelle .....	68

---

Figure III.3. Evaluation objective prise en compte lors de la phase de data mining dans notre démarche.....	70
Figure III.4. Illustration des limites des mesures classiques Support-Confiance.....	70
Figure III.5. Evaluation sémantique dans la phase de post-traitement .....	73
Figure III.6. Liens entre les différentes règles extraites ( <i>sup, conf</i> ).....	79
Figure III.7. Métarègle regroupant les items des règles extraites et leurs liens .....	81
Figure III.8. Illustration de l'opération de projection de la règle R12 dans la métarègle .....	81
Figure III.9. Illustration d'une requête ( <i>Q</i> ) de l'utilisateur exprimée par un GC.....	84
Figure III.10. Illustration d'une règle conforme à la requête de l'utilisateur .....	85
Figure III.11. . Illustration de règles avec partie conclusion inattendue .....	86
Figure III.12. Illustration de règles avec partie hypothèse inattendue .....	86
Figure III.13. Illustration de règles inattendues par l'utilisateur .....	87
Figure IV.1. Types de maintenance selon la norme AFNOR NF EN 13306 .....	91
Figure IV.2. Rapports des interventions de maintenance (Cas # 1) .....	97
Figure IV.3. Rapports des interventions de maintenance (Cas # 2) .....	97
Figure IV.4. Rapports des interventions de maintenance (Cas # 3) .....	98
Figure IV.5. Interface de l'outil CoGui editor.....	99
Figure IV.6. Visualisation d'une partie de la hiérarchie des types de concepts ( <i>TC</i> ) modélisées dans CoGui (Cas # 1) .....	100
Figure IV.7. Hiérarchie des types de relation ( <i>TR</i> ) modélisées dans CoGui .....	101
Figure IV.8. Modèle générique d'une expérience (Cas # 1) .....	102
Figure IV.9. Illustration d'une intervention du cas #1 .....	103
Figure IV.10. Interface du logiciel SPMF.....	106
Figure IV.11. Evolution des résultats obtenus lorsque le <i>minsup</i> et le <i>minconf</i> varient (Cas # 1) ...	107
Figure IV.12. Evolution des résultats obtenus lorsque le <i>minsup</i> et le <i>minconf</i> varient (Cas # 2) ...	109
Figure IV.13. Liens entre les règles extraites ( <i>sup, conf</i> ) - Illustration du cas # 1 .....	126
Figure IV.14. Métarègle regroupant tous les items des règles extraites et leurs liens (Cas #1) .....	127
Figure IV.15. Illustration la règle R10 contenue dans la métarègle .....	127
Figure IV.16. Requête ( <i>Q</i> ) de l'utilisateur .....	128
Figure IV.17. Règle R8 avec partie conclusion inattendue .....	129
Figure IV.18. Règle R1 avec partie hypothèse inattendue .....	129
Figure IV.19. Règle R15 avec partie hypothèse inattendue .....	130
Figure IV.20. Règle R3 avec hypothèse et conclusion inattendues.....	130

---

## LISTE DE TABLEAUX

Tableau III.1. Ensemble de règles extraites satisfaisant un <i>minsup</i> = 30% et un <i>minconf</i> = 50% .....	75
Tableau IV.1. Revue des études liées à l'extraction de règles d'association en maintenance .....	94
Tableau IV.2. Illustration du contexte formel (Cas # 1).....	105
Tableau IV.3. Illustration du contexte formel (Cas # 2).....	105
Tableau IV.4. Résultats obtenus - 1067 interventions et 12 attributs (Cas # 1) .....	107
Tableau IV.5. Résultats obtenus - 5955 interventions et 15 attributs (Cas # 2) .....	108
Tableau IV.6. Résultats obtenus - 537 interventions et 16 attributs (Cas # 3) .....	109
Tableau IV.7. Règles d'association extraites ( <i>minsup</i> = 30%, <i>minconf</i> = 90%) - Cas # 1 .....	110
Tableau IV.8. Questions associées aux attributs présents - Cas # 1 .....	111
Tableau IV.9. Règles d'association ( <i>minsup</i> = 50%, <i>minconf</i> = 90%) - Cas #2 (Test 1) .....	113
Tableau IV.10. Nouveaux résultats obtenus - 5955 interventions et 10 attributs (Cas # 2) .....	115
Tableau IV.11. Règles d'association ( <i>minsup</i> = 20%, <i>minconf</i> = 90%) - Cas #2 (Test 2) .....	116
Tableau IV.12. Règles d'association ( <i>minsup</i> = 40%, <i>minconf</i> = 100%) - Cas # 3 (Test 1) .....	117
Tableau IV.13. Questions associées aux attributs présents - Cas # 3 (Test 1) .....	117
Tableau IV.14. Nouveaux résultats obtenus - 537 interventions et 12 attributs (Cas # 3) .....	118
Tableau IV.15. Règles d'association ( <i>minsup</i> = 10%, <i>minconf</i> = 70%) - Cas # 3 (Test 2) .....	119
Tableau IV.16. Questions associées aux attributs présents- Cas # 3 (Test 2) .....	119



## ALGORITHMES

Algorithme I.1. Découverte des <i>itemsets</i> fréquents dans l'algorithme <i>Apriori</i> .....	30
Algorithme I.2. Génération des règles dans l'algorithme <i>Apriori</i> (Agrawal & Srikant, 1994) .....	31



## BIBLIOGRAPHIE

- Aamodt, A. & Plaza, E. (1994), 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AI communications* **7**(1), 39–52.
- Agarwal, R. C., Aggarwal, C. C. & Prasad, V. V. V. (2001), 'A tree projection algorithm for generation of frequent item sets', *Journal of Parallel and Distributed Computing* **61**(3), 350–371.
- Agrawal, R., Imielinski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, in 'Proceedings of the ACM SIGMOD International Conference on Management of Data', SIGMOD'93, ACM, Washington, D.C., USA, 26-28 mai, pp. 207–216.
- Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules in large databases, in 'Proceedings of the 20th International Conference on Very Large Data Bases', Vol. 1215 of *VLDB'94*, Morgan Kaufmann Publishers Inc., 12-15 sept., pp. 487–499.
- Agrawal, R. & Srikant, R. (1995), Mining sequential patterns, in 'Proceedings of the 11th International Conference on Data Engineering', ICDE '95, IEEE Computer Society, Washington, DC, USA, 6-10 mars, pp. 3–14.
- Alsyouf, I. (2009), 'Maintenance practices in swedish industries: Survey results', *International Journal of Production Economics* **121**(1), 212–223.
- Alzghoul, A. & Löfstrand, M. (2011), 'Increasing availability of industrial systems through data stream mining', *Computers & Industrial Engineering* **60**(2), 195–205.
- Apedome, K. S., Addouche, S. & EL Mhamedi, A. (2011), 'L'intégration des aspects cognitifs au retour d'expérience statistique dans la résolution de problèmes industriels'.
- Armaghan, N. (2009), Contribution à un système de retour d'expérience basé sur le raisonnement à partir de cas conversationnel : application à la gestion des pannes de machines industrielles, Thèse de doctorat, Institut National Polytechnique de Lorraine, France.
- Ashrafi, M. Z., Taniar, D. & Smith, K. (2007), 'Redundant association rules reduction techniques', *International Journal of Business Intelligence and Data Mining* **2**(1), 29–63.
- Azé, J. (2003), Extraction de connaissances à partir de données numériques et textuelles, Thèse de doctorat, Université Paris-Sud, France.
- Azé, J. & Kodratoff, Y. (2002), Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association, in D. H. et Djamel A. Zighed, ed., 'Extraction et gestion des connaissances EGC', Vol. 1 of *Extraction des connaissances et apprentissage*, Hermes Science Publications, pp. 143–154.

- 
- Baesens, B., Viaene, S. & Vanthienen, J. (2000), 'Post-processing of association rules', *DTEW Research Report 0020* pp. 1–18.
- Baget, J. F. (2001), Représenter des connaissances et raisonner avec des hypergraphes : de la projection à la dérivation sous contraintes, Thèse de doctorat, Université de Montpellier II - Sciences et Techniques du Languedoc, France.
- Baget, J. F., Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., Gandon, F., Giboin, A., Gutierrez, A., Leclère, M., Mugnier, M. L. & Thomopoulos, R. (2008), Griwes: Generic model and preliminary specifications for a graph-based knowledge representation toolkit, in P. Eklund & O. Haemmerle, eds, 'Conceptual Structures: Knowledge Visualization and Reasoning', Vol. 5113 of *Lecture Notes in Artificial Intelligence*, Springer Berlin Heidelberg, pp. 297–310.
- Baget, J. F. & Mugnier, M. L. (2002), 'Extension of simple conceptual graphs: the complexity of rules and constraints', *Journal of Artificial Intelligence Research* **16**, 425–465.
- Baohui, J., Yuxin, W. & Zheng-qing, Y. (2011), The research of data mining in ahm technology based on association rule, in 'Prognostics and System Health Management Conference (PHM)', IEEE, Shenzhen, Chine, 24-25 mai, pp. 1–8.
- Bastide, Y., Taouil, R., Pasquier, N., Stumme, G. & Lakhal, L. (2000), 'Mining frequent patterns with counting inference', *ACM SIGKDD Explorations Newsletter* **2**(2), 66–75.
- Batanov, D., Nagarur, N. & Nitikhunkasem, P. (1993), 'Expert-mm: A knowledge-based system for maintenance management', *Artificial Intelligence in Engineering* **8**(4), 283–291.
- Ben Ayed, M., Ltifi, H., Kolski, C. & Alimi, A. M. (2010), 'A user-centered approach for the design and implementation of kdd-based dss: A case study in the healthcare domain', *Decision Support Systems* **50**(1), 64–78.
- Ben-Daya, M., Duffuaa, S. O., Raouf, A., Knezevic, J. & Ait-Kadi, D. (2009), *Handbook of maintenance management and engineering*, Vol. 1, Springer.
- Ben Sta, H. (2006), Contribution de la modelisation conceptuelle à l'ingénierie du knowledge management : Application dans le cadre de la mémoire de projet, Thèse de doctorat, Ecole Centrale de Lille et Ecole Centrale de Paris, France.
- Bendoly, E. (2003), 'Theory and support for process frameworks of knowledge discovery and data mining from erp systems', *Information & Management* **40**(7), 639–647.
- Benomrane, S., Ben Ayed, M. & Alimi, A. M. (2013), An agent-based knowledge discovery from databases applied in healthcare domain, in 'International Conference on Advanced Logistics and Transport', ICALT 2013, Sousse, Tunisie, 29-31 mai, pp. 176–180.
- Bergmann, R. (2002), *Experience management: foundations, development methodology, and internet-based applications*, Vol. 2432 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin, Heidelberg.
- Bertin, A. (2012), Intégration d'un système de retour d'expériences à un PLM, Thèse de doctorat, Institut National Polytechnique de Toulouse (INPT), France.
- Bickford, J. C. (2000), Sharing lessons learned in the department of energy, in 'AAAI-00 Intelligent Lessons Learned Systems Workshop', Austin, Texas, USA, 31 juillet, pp. 5–8.



- Blanchard, J. (2005), Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association, Thèse de doctorat, Université de Nantes, France.
- Béler, C. (2008), Modélisation générique d'un retour d'expérience cognitif : Application à la prévention des risques, Thèse de doctorat, Institut National Polytechnique de Toulouse (INPT), France.
- Borgida, A. (1996), 'On the relative expressiveness of description logics and predicate logics', *Artificial Intelligence* **82**(1/2), 353–367.
- Botta-Genoulaz, V., Millet, R. & Grabot, B. (2005), 'A survey on the recent research literature on erp systems', *Computers in Industry* **56**(6), 510–522.
- Breuker, J. (2013), 'A cognitive science perspective on knowledge acquisition', *International Journal of Human-Computer Studies* **71**(2), 177–183.
- Brin, S., Motwani, R. & Silverstein, C. (1997), Beyond market baskets: Generalizing association rules to correlations, in J. Peckham, ed., 'Proceedings of the ACM SIGMOD International Conference on Management of Data', SIGMOD '97, ACM Press, Tucson, Arizona, USA, 13-15 mai, pp. 265–276.
- Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. (1997), Dynamic itemset counting and implication rules for market basket data, in J. Peckham, ed., 'Proceedings of the ACM SIGMOD International Conference on Management of Data', SIGMOD '97, ACM Press, Tucson, Arizona, USA, 13-15 mai, pp. 255–264.
- Cao, L. & Zhang, C. (2008), Domain driven data mining, in D. Tanar, ed., 'Data Mining and Knowledge Discovery Technologies', IGI Global, Hershey, New York, USA, chapter IX, pp. 196–223.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000), *CRISP-DM 1.0 Step-by-step data mining guide*.
- Chassiakos, A. P., Vagiotas, P. & Theodorakopoulos, D. D. (2005), 'A knowledge-based system for maintenance planning of highway concrete bridges', *Advances in Engineering Software* **36**(11-12), 740–749.
- Chebel-Morello, B. (2008), Définition d'un modèle générique des systèmes de retour d'expérience en entreprise, in '3ème Conférence Francophone de Modélisation et Simulation "Conception, Analyse et Gestion des Systèmes Industriels"', MOSIM'08, Paris, France, 31 mars - 2 avril.
- Chein, M. & Mugnier, M. L. (1992), 'Conceptual graphs: Fundamental notions', *Revue d'Intelligence Artificielle* **6**(4), 365–406.
- Chein, M. & Mugnier, M. L. (2008), *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*, Advanced Information and Knowledge Processing, Springer Publishing Company, Incorporated, London, UK.
- Chein, M., Mugnier, M. L. & Croitoru, M. (2013), 'Visual reasoning with graph-based mechanisms: the good, the better and the best', *The Knowledge Engineering Review* **28**(3), 249–271.
- Chen, C. L. P. & Zhang, C. Y. (2014), 'Data-intensive applications, challenges, techniques and technologies: A survey on big data', *Information Sciences* **275**, 314–347.
- Chen, E., Cao, H., Li, Q. & Qian, T. (2008), 'Efficient strategies for tough aggregate constraint-based sequential pattern mining', *Information Sciences* **178**(6), 1498 – 1518.

- 
- Chen, M. C. (2003), 'Configuration of cellular manufacturing systems using association rule induction', *International Journal of Production Research* **41**(2), 381–395.
- Chen, M. S., Han, J. & Yu, P. S. (1996), 'Data mining: an overview from a database perspective', *IEEE Transactions on Knowledge and Data Engineering* **8**(6), 866–883.
- Chen, W. C., Tseng, S. S. & Wang, C. Y. (2005), 'A novel manufacturing defect detection method using association rule mining techniques', *Expert Systems with Applications* **29**(4), 807–815.
- Chen, Y. J. (2010), 'Development of a method for ontology-based empirical knowledge representation and reasoning', *Decision Support Systems* **50**(1), 1–20.
- Chevrin, V., Couturier, O., Mephu Nguifo, E. & Rouillard, J. (2007), 'Recherche anthropocentrée de règles d'association pour l'aide à la décision', *Revue d'Interaction Homme-Machine* **8**(2).
- Chiang, F. & Miller, R. J. (2008), 'Discovering data quality rules', *Proceedings of the VLDB Endowment* **1**(1), 1166–1177.
- Choo, C. W. (1998), *The knowing organization: How organizations use information to construct meaning, create knowledge, and make decisions*, Vol. 256, Oxford University Press, New York.
- Choudhary, A. K., Harding, J. A. & Tiwari, M. K. (2009), 'Data mining in manufacturing: a review based on the kind of knowledge', *Journal of Intelligent Manufacturing* **20**(5), 501–521.
- Clermont, P., Béler, C., Rakoto, H., Desforges, X. & Geneste, L. (2007), Capitalisation et exploitation du retour d'expérience : un raisonnement à partir de cas étendu aux systèmes sociotechniques, in J. Renaud, B. Chebel-Morello, B. Fuchs & J. Lieber, eds, 'Raisonnement à partir de cas 1 : conception et configuration de produits', Hermès - Lavoisier, chapter 8, pp. 249–277.
- Cooke, P. & Leydesdorff, L. (2006), 'Regional development in the knowledge-based economy: The construction of advantage', *The Journal of Technology Transfer* **31**(1), 5–15.
- Couturier, O. (2005), Contribution à la fouille de données : règles d'association et interactivité au sein d'un processus d'extraction de connaissances dans les données, Thèse de doctorat, Université d'Artois, France.
- Crespo Marquez, A. & Gupta, J. N. D. (2006), 'Contemporary maintenance management: process, framework and supporting pillars', *Omega - International Journal of Management Science* **34**(3), 313–326.
- Dasu, T. & Johnson, T. (2003), *Exploratory data mining and data cleaning*, Vol. 479, John Wiley & Sons, Inc, New York, NY.
- Dau, F. & Eklund, P. (2008), 'A diagrammatic reasoning system for the description logic alc', *Journal of Visual Languages and Computing* **19**(5), 539–573.
- Delange, L. & Vogin, R. (1994), 'La croissance de sûreté de fonctionnement par le retour d'expérience dans le domaine technique et industriel', *Performances Humaines et Techniques* **69**.
- Delgado, M., Sanchez, D., Martin-Bautista, M. J. & Vila, M. A. (2001), 'Mining association rules with improved semantics in medical databases', *Artificial Intelligence in Medicine* **21**(1-3), 241–245.
- Dienes, Z. & Perner, J. (1999), 'A theory of implicit and explicit knowledge', *Behavioral and Brain Sciences* **22**(5), 735–808.

- Dieng-Kuntz, R. & Corby, O. (2005), Conceptual graphs for semantic web applications, in F. Dau, M. L. Mugnier & G. Stumme, eds, 'Conceptual Structures: Common Semantics for Sharing Knowledge', Vol. 3596 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, pp. 19–50.
- Dieng-Kuntz, R., Corby, O., Gandon, F., Giboin, A., Golebiowska, J., Matta, N. & Ribiere, M. (2001), *Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du knowledge management*, Vol. 2 of *Systèmes d'Information*, Dunod Edition Informatiques.
- Dorn, M., Hou, W., Che, D. & Jiang, Z. (2008), 'An empirical study of qualities of association rules from a statistical view point', *Journal of Information Processing Systems* **4**(1), 27–32.
- EN13306:2001 (2001), *Maintenance Terminology - European Standard*. CEN (European Committee for Standardization), Brussels.
- Ermine, J. L. (2000), *Les systèmes de connaissances*, Vol. 2ème edition, Hermès sciences publication, Paris.
- Ermine, J. L., Chaillot, M., Bigeon, P., Charreton, B. & Malavieille, D. (1996), 'Mksm : Méthode pour la gestion des connaissances', *Ingénierie des systèmes d'information* **4**(4), 541–575.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**(3), 37–54.
- Fournier-Viger, P., Wu, C. W. & Tseng, V. S. (2012), Mining top-k association rules, in L. Kosseim & D. Inkpen, eds, 'Advances in Artificial Intelligence', Vol. 7310 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 61–73.
- Frawley, W. J., Piatetsky-Shapiro, G. & Matheus, C. J. (1992), 'Knowledge discovery in databases - an overview', *AI Magazine* **13**(3), 57–70.
- Fürst, F., Leclère, M. & Trichet, F. (2003), 'Ontological engineering and mathematical knowledge management: A formalization of projective geometry', *Annals of Mathematics and Artificial Intelligence* **38**(1-3), 65–89.
- Fürst, F. & Trichet, F. (2009), 'Axiom-based ontology matching', *Expert Systems* **26**(2), 218–246.
- Garcia, E., Romero, C., Ventura, S. & de Castro, C. (2009), 'An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering', *User Modeling and User-Adapted Interaction* **19**(1-2), 99–132.
- Gasmi, G., Ben Yahia, S., Mephu Nguifo, E. & Slimani, Y. (2006), 'lgb: une nouvelle base générique informative des règles d'association', *Revue I3 (Information Interaction Intelligence)* **6**(1), 31–67.
- Geng, L. & Hamilton, H. J. (2006), 'Interestingness measures for data mining: A survey', *ACM Computing Surveys* **38**(3), Article No. 9.
- Giudici, P. (2003), *Applied data mining: Statistical methods for business and industry*, John Wiley & Sons Ltd.
- Goodman, R. M. & Smyth, P. (1988), Information-theoretic rule induction, in 'ECAI', pp. 357–362.
- Gras, R. (1979), Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse d'état, Université de Rennes I, France.

- 
- Gras, R., Kuntz, P. & Briand, H. (2001), 'Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données', *Mathématiques et Sciences Humaines* **154-155**, 9–29.
- Gruber, T. R. (1995), 'Toward principles for the design of ontologies used for knowledge sharing', *International Journal of Human-Computer Studies* **43**(5-6), 907–928.
- Grundstein, M. (2004), De la capitalisation des connaissances au management des connaissances dans l'entreprise, in I. Boughzala & J. L. Ermine, eds, 'Management des connaissances en entreprise', Edition Hermès Science, Paris, chapter 1, pp. 25–54.
- Han, J. & Kamber, M. (2001), 'Data mining: concept and technology', *Publishing House of Mechanism Industry* pp. 70–72.
- Han, J. & Kamber, M. (2006), *Data mining: concepts and techniques*, Vol. 2, Morgan Kaufmann Publishers, San Francisco, CA.
- Han, J. & Pei, J. (2000), 'Mining frequent patterns by pattern-growth: Methodology and implications', *ACM SIGKDD Explorations Newsletter* **2**(2), 14–20.
- Harding, J. A., Shahbaz, M., Srinivas & Kusiak, A. (2005), 'Data mining in manufacturing: A review', *Journal of Manufacturing Science and Engineering - transactions of the ASME* **128**(4), 969–976.
- Hell, P. & Nesetril, J. (2004), *Graphs and Homomorphisms*, Vol. 28, Oxford Lecture Series in Mathematics and Its Applications.
- Hellerstein, J. M. (2008), 'Quantitative data cleaning for large databases', *United Nations Economic Commission for Europe (UNECE)* .
- Hermiz, K. B. (1999), 'Critical success factors for data mining projects', *DM Review Magazine* .
- Hogan, J., Hardiman, F. & Naughton, M. D. (2011), Asset management: A review of contemporary & individualised strategies, in 'Proceedings of the World Congress on Engineering WCE', Vol. 1 of *Lecture notes in engineering and computer science*, Londres, 6-8 juillet, pp. 545–549.
- Houtsma, M. & Swami, A. (1993), Set-oriented mining of association rules, Research report rj 9567, IBM Almaden Research Center, San Jose, California.
- Huang, J. J., Tzeng, G. H. & Ong, C. S. (2007), 'Marketing segmentation using support vector clustering', *Expert Systems with Applications* **32**(2), 313–317.
- Huang, R., Xi, L., Lee, J. & Liu, C. R. (2005), 'The framework, impact and commercial prospects of a new predictive maintenance system: intelligent maintenance system', *Production Planning & Control* **16**(7), 652–664.
- International Business Machines, I. (1996), Ibm intelligent miner user's guide, Technical Report Version 1, Release 1, SH12-6213-00 edition, juillet.
- Jabrouni, H. (2012), Exploitation des connaissances issues des processus de retour d'expérience industriels, Thèse de doctorat, Institut National Polytechnique de Toulouse (INPT), France.
- Jiao, J., Zhang, L., Zhang, Y. & Pokharel, S. (2008), 'Association rule mining for product and process variety mapping', *International Journal of Computer Integrated Manufacturing* **21**(1), 111–124.

- Kamsu-Foguem, B., Coudert, T., Béler, C. & Geneste, L. (2008), 'Knowledge formalization in experience feedback processes: An ontology-based approach', *Computers in Industry* **59**(7), 694–710.
- Kamsu-Foguem, B. & Mathieu, Y. (2014), 'Software architecture knowledge for intelligent light maintenance', *Advances in Engineering Software* **67**, 125–135.
- Kamsu-Foguem, B. & Noyes, D. (2013), 'Graph-based reasoning in collaborative knowledge management for industrial maintenance', *Computers in Industry* **64**(8), 998–1013.
- Kamsu-Foguem, B., Rigal, F. & Mauget, F. (2013), 'Mining association rules for the quality improvement of the production process', *Expert Systems with Applications* **40**(4), 1034–1045.
- Kargupta, H., Han, J., Yu, P. S., Motwani, R. & Kumar, V. (2008), *Next Generation of Data Mining, Data Mining and Knowledge Discovery*, 1 edn, Chapman & Hall/CRC.
- Karray, M. H., Chebel-Morello, B. & Zerhouni, N. (2014), 'Petra: Process evolution using a trace-based system on a maintenance platform', *Knowledge-Based Systems* **68**, 21–39.
- Kasravi, K. (1997), Data mining and knowledge discovery in manufacturing, in 'AUTOFACT'97, Information Technologies for the Manufacturing Enterprise', Michigan.
- Kimble, C. & Bourdon, I. (2008), 'Some success factors for the communal management of knowledge', *International Journal of Information Management* **28**(6), 461–467.
- Köksal, G., Batmaz, I. & Testik, M. C. (2011), 'A review of data mining applications for quality improvement in manufacturing industry', *Expert Systems with Applications* **38**(10), 13448–13467.
- Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. & Verkamo, A. I. (1994), Finding interesting rules from large sets of discovered association rules, in 'Proceedings of the third International Conference on Information and Knowledge Management (CIKM)', ACM, Gaithersburg, MD, USA, 29 nov. - 2 déc., pp. 401–407.
- Kolb, D. (2000), The process of experiential learning, in R. L. Cross & S. B. Israelit, eds, 'Strategic Learning in a Knowledge Economy', Butterworth-Heinemann, chapter 15, pp. 313–331.
- Koskinen, K. U. (2012), 'Problem absorption as an organizational learning mechanism in project-based companies: Process thinking perspective', *International Journal of Project Management* **30**(3), 308–316.
- Kuntz, P., Guillet, F., Lehn, R. & Briand, H. (2000), A user-driven process for mining association rules, in D. A. Zighed, J. Komorowski & J. Zykwow, eds, 'Principles of Data Mining and Knowledge Discovery', Vol. 1910 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 483–489.
- Kusiak, A. (2002), 'A data mining approach for generation of control signatures', *Journal of Manufacturing Science and Engineering* **124**(4), 923–926.
- Lacolare, V. (2011), *Valoriser la connaissance dans l'entreprise*, Solutions pour, Afnor edition.
- Lai, L. F. (2007), 'A knowledge engineering approach to knowledge management', *Information Sciences* **177**(19), 4072–4094.
- Lallich, S. & Teytaud, O. (2004), 'Évaluation et validation de mesures d'intérêt des règles d'association', *Revue des Nouvelles Technologies de l'Information RNTI-E-1* pp. 193–218.



- 
- Larose, D. T. (2005), *Discovering knowledge in data: an introduction to data mining*, Wiley-Interscience.
- Lavrac, N., Flach, P. & Zupan, B. (1999), Rule evaluation measures: A unifying view, in S. Dzeroski & P. Flach, eds, 'Inductive Logic Programming', Vol. 1634 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 174–185.
- Lazcorreta, E., Botella, F. & Fernandez-Caballero, A. (2008), 'Towards personalized recommendation by two-step modified apriori data mining algorithm', *Expert Systems with Applications* **35**(3), 1422–1429.
- Lee, D. H. & Kim, M. (1997), 'Database summarization using fuzzy isa hierarchies', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **27**(1), 68–78.
- Lejarraga, T. & Gonzalez, C. (2011), 'Effects of feedback and complexity on repeated decisions from description', *Organizational Behavior and Human Decision Processes* **116**(2), 286–295.
- Lerman, I. C. (1981), *Classification et analyse ordinale des donnees*, Dunod, Paris.
- Lerman, I. C., Gras, R. & Rostam, H. (1981), 'Élaboration et évaluation d'un indice d'implication pour des données binaires', *Mathématiques et Sciences Humaines* **75**, 5–47.
- Lesser, E. & Prusak, L. (2004), *Creating value with knowledge: Insights from the IBM Institute for business value*, Oxford University Press.
- Liao, S. H. (2003), 'Knowledge management technologies and applications - literature review from 1995 to 2002', *Expert Systems with Applications* **25**(2), 155–164.
- Liao, S. H., Hsieh, C. L. & Huang, S. P. (2008), 'Mining product maps for new product development', *Expert Systems with Applications* **34**(1), 50–62.
- Lieber, J. (2007), Application of the revision theory to adaptation in case-based reasoning: the conservative adaptation, in R. O. Weber & M. M. Richter, eds, 'Proceedings of Case-Based Reasoning Research and Development', Vol. 4626 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 239–253.
- Liebowitz, J. (2001), 'Knowledge management and its link to artificial intelligence', *Expert Systems with Applications* **20**(1), 1–6.
- Liu, B., Hsu, W. & Chen, S. (1997), Using general impressions to analyze discovered classification rules, in 'Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)', AAAI Press, pp. 31–36.
- Liu, B., Hsu, W., Wang, K. & Chen, S. (1999), Visually aided exploration of interesting association rules, in N. Zhong & L. Zhou, eds, 'Methodologies for Knowledge Discovery and Data Mining', Vol. 1574 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 380–389.
- Liu, H. & Motoda, H. (1998), *Feature Selection for Knowledge Discovery and Data Mining*, Vol. 545 of *The Springer International Series in Engineering and Computer Science*, Kluwer Academic Publishers, Norwell, MA, USA.
- Liu, Y., Xu, W. & Du, H. (2011), The method of test for state of railway tunnel lining based on association rules, in 'Proceedings of the 3rd International Conference on Communication Software and Networks (ICCSN)', IEEE, Xi'an, Chine, 27-29 mai, pp. 387–390.

- Létourneau, S., Famili, F. & Matwin, S. (1999), 'Data mining to predict aircraft component replacement', *IEEE Intelligent Systems & their Applications* **14**(6), 59–66.
- Lu, J. & Ruan, D. (2007), 'Intelligent knowledge engineering systems', *Knowledge-Based Systems* **20**(5), 437–438.
- Lœvinger, J. (1947), 'A systematic approach to the construction and evaluation of tests of ability', *Psychological Monographs* **61**(4), 1–49.
- Maquee, A., Shojaie, A. A. & Mosaddar, D. (2012), 'Clustering and association rules in analyzing the efficiency of maintenance system of an urban bus network', *International Journal of System Assurance Engineering and Management* **3**(3), 175–183.
- Marban, O., Segovia, J., Menasalvas, E. & Fernandez-Baizan, C. (2009), 'Toward data mining engineering: A software engineering approach', *Information Systems* **34**(1), 87–107.
- Marinica, C. (2010), Association Rule Interactive Post-processing using Rule Schemas and Ontologies-ARIPSO, Thèse de doctorat, Ecole Polytechnique de l'Université de Nantes, France.
- Matta, N., Ermine, J. L., Aubertin, G. & Trivin, J. Y. (2002), Knowledge capitalization with a knowledge engineering approach: The mask method, in R. Dieng-Kuntz & N. Matta, eds, 'Knowledge Management and Organizational Memories', Part I, Springer US, pp. 17–28.
- McGuinness, D. L. & Borgida, A. T. (1995), Explaining subsumption in description logics, in 'Proceedings of the 14th International Joint Conference on Artificial Intelligence', IJCAI'95, Morgan Kaufmann Publishers Inc., Montréal, Canada, 20-25 août, pp. 816–821.
- Meseroll, R. J., Kirkos, C. J. & Shannon, R. A. (2007), Data mining navy flight and maintenance data to affect repair, in '42nd Annual AUTOTESTCON Conference', Vol. 1 and 2, IEEE, Baltimore, MD, USA, 17-20 sept, pp. 476–481.
- Millot, P. & Debernard, S. (1993), Men-machines cooperative organizations: methodological and practical attempts in air traffic control, in 'Proceedings of the International Conference on Systems, Man and Cybernetics', Vol. 1, IEEE, Le Touquet, France, 17-20 oct., pp. 695–700.
- Minor, M. (2005), Introduction strategy and feedback from an experience management project, in K. D. Althoff, A. Dengel, R. Bergmann, M. Nick & T. Roth-Berghofer, eds, 'Professional Knowledge Management', Vol. 3782 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 284–292.
- Minsky, M. (1975), A framework for representing knowledge, in P. Winston, ed., 'The Psychology of Computer Vision', McGraw-Hill, pp. 211–277.
- Mirabadi, A. & Sharifian, S. (2010), 'Application of association rules in iranian railways (rai) accident data analysis', *Safety Science* **48**(10), 1427–1435.
- Mom, T. J. M., Van Den Bosch, F. A. J. & Volberda, H. W. (2007), 'Investigating managers' exploration and exploitation activities: The influence of top-down, bottom-up, and horizontal knowledge inflows\*', *Journal of Management Studies* **44**(6), 910–931.
- Morita, T., Satou, Y., Ayukawa, E. & Maeda, A. (2000), Customer relationship management through data mining, in 'Informs-Korms Seoul 2000', pp. 1956–1963.
- Moubray, J. (1991), *Reliability Centred Maintenance*, Butterworth-Heinemann.

- 
- Mugnier, M. L. (1995), 'On generalization/specialization for conceptual graphs', *Journal of Experimental & Theoretical Artificial Intelligence* **7**(3), 325–344.
- Mugnier, M. L. (2000), Knowledge representation and reasonings based on graph homomorphism, in B. Ganter & G. W. Mineau, eds, 'Proceedings of Conceptual Structures: Logical, Linguistic, and Computational Issues', Vol. 1867 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 172–192.
- Musen, M. A. (1993), An overview of knowledge acquisition, in J. M. David, J. P. Krivine & R. Simmons, eds, 'Second Generation Expert Systems', Part IV, Springer Berlin Heidelberg, pp. 405–427.
- Nahar, J., Imam, T., Tickle, K. S. & Chen, Y. P. P. (2013), 'Association rule mining to detect factors which contribute to heart disease in males and females', *Expert Systems with Applications* **40**(4), 1086–1093.
- Nakagawa, T. (1986), 'Periodic and sequential preventive maintenance policies', *Journal of Applied Probability* **23**(2), 536–542.
- Nakajima, S. (1988), 'Introduction to tpm: Total productive maintenance', Productivity Press, Cambridge, Mass. (translated into English from the original text published by the Japan Institute for Plant Maintenance, 1984).
- Nemati, H. R., Steiger, D. M., Iyer, L. S. & Herschel, R. T. (2002), 'Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing', *Decision Support Systems* **33**(2), 143–161.
- Ngai, E. W. T., Xiu, L. & Chau, D. C. K. (2009), 'Application of data mining techniques in customer relationship management: A literature review and classification', *Expert Systems with Applications* **36**(2), 2592–2602.
- Nisbet, R., Elder, J. & Miner, G. (2009), *Handbook of statistical analysis and data mining applications*, Academic Press.
- Nonaka, I. & Takeuchi, H. (1995), *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, Inc.
- Nonaka, I., Umemoto, K. & Senoo, D. (1996), 'From information processing to knowledge creation: A paradigm shift in business management', *Technology in Society* **18**(2), 203–218.
- Park, J. S., Chen, M. S. & Yu, P. S. (1997), 'Using a hash-based method with transaction trimming for mining association rules', *IEEE Transactions on Knowledge and Data Engineering* **9**(5), 813–825.
- Pasquier, N. (2000), Data Mining : Algorithmes d'Extraction et de Réduction des Règles d'Association dans les Bases de Données, Thèse de doctorat, Université Clermont-Ferrand II, France.
- Pasquier, N., Bastide, Y., Taouil, R. & Lakhal, L. (1999), 'Efficient mining of association rules using closed itemset lattices', *Information Systems* **24**(1), 25–46.
- Pasquier, N., Taouil, R., Bastide, Y., Stumme, G. & Lakhal, L. (2005), 'Generating a condensed representation for association rules', *Journal of Intelligent Information Systems* **24**(1), 29–60.
- Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, number 552, Morgan Kaufmann.



- Pechenizkiy, M., Puuronen, S. & Tsymbal, A. (2005), Why data mining research does not contribute to business?, in C. Soares, L. Moniz & C. Duarte, eds, 'Proceedings of Data Mining for Business Workshop at ECLM/PKDD Conference', DMbiz'05, Oxford University Press, Porto, Portugal, 3-7 oct., pp. 67–71.
- Pechenizkiy, M., Puuronen, S. & Tsymbal, A. (2008), 'Towards more relevance-oriented data mining research', *Intelligent Data Analysis - Philosophies and Methodologies for Knowledge* **12**(2), 237–249.
- Pei, J., Han, J. & Mao, R. (2000), Closet: An efficient algorithm for mining frequent closed itemsets, in D. Gunopulos & R. Rastogi, eds, 'Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery', Vol. 4 of *DMKD'00*, Dallas, Texas, USA, 14 mai, pp. 21–30.
- Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI Press, pp. 229–248.
- Pieri, G., Klein, M. R. & Milanese, M. (2002), 'Maic: A data and knowledge-based system for supporting the maintenance of chemical plant', *International Journal of Production Economics* **79**(2), 143–159.
- Poncelet, P., Massegia, F. & Teisseire, M. (2008), *Data Mining Patterns: New Methods and Applications*, IGI Global, Hershey, PA.
- Potes Ruiz, P. A., Kamsu-Foguem, B. & Noyes, D. (2013), 'Knowledge reuse integrating the collaboration from experts in industrial maintenance management', *Knowledge-Based Systems* **50**, 171–186.
- Potes Ruiz, P. A., Noyes, D. & Kamsu-Foguem, B. (2012), Raisonement collaboratif à partir de cas dans la résolution de problèmes en maintenance, in 'Proceedings of the 9th International Conference on Modeling, Optimization & SIMulation', MOSIM'12, 6-8 juin, Bordeaux, France.
- Quan, T., Ngo, L. & Hui, S. (2009), An effective clustering-based approach for conceptual association rules mining, in 'Proceedings IEEE International Conference on Computing and Communication Technologies', IEEE Computer Society, Washington, D.C., pp. 1–7.
- Quillian, M. R. (1968), Semantic memory, in M. Minsky, ed., 'Semantic Information Processing', MIT Press, pp. 227–270.
- Rachidi, A., Talbi, A. & Khatory, A. (2013), 'The new forms of the industrial maintenance: Which impact in the performance of the industrial companies? (case study)', *International Journal of Engineering and Advanced Technology (IJEAT)* **2**(5), 500–507.
- Rahm, E. & Do, H. H. (2000), 'Data cleaning: Problems and current approaches', *IEEE Bulletin on Data Engineering* **23**(4), 3–13.
- Rakoto, H. (2004), Intégration du retour d'expérience dans les processus industriels : application à Alstom Transport, Thèse de doctorat, Institut National Polytechnique de Toulouse (INPT), France.
- Rakoto, H., Clermont, P. & Geneste, L. (2002), Elaboration and exploitation of lessons learned, in M. A. Musen, B. Neumann & R. Studer, eds, 'Intelligent Information Processing', Vol. 93, Springer US, pp. 297–300.

- 
- Rasovska, I., Chebel-Morello, B. & Zerhouni, N. (2008), 'A mix method of knowledge capitalization in maintenance', *Journal of Intelligent Manufacturing* **19**(3), 347–359.
- Renaud, J., Bonjour, E., Chebel-Morello, B., Fuchs, B. & Matta, N. (2008), *Retour et capitalisation d'expérience - Outils et démarches*, AFNOR Editions, Paris.
- Rexao (2003), Groupement rexao (retour d'expérience et apprentissage organisationnel), Rapport technique.
- Rosqvist, T., Laakso, K. & Reunanen, M. (2009), 'Value-driven maintenance planning for a production plant', *Reliability Engineering & System Safety* **94**(1), 97–110.
- Sahar, S. (1999), Interestingness via what is not interesting, in U. Fayyad, S. Chaudhuri & D. Madigan, eds, 'Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '99, San Diego, CA, USA, 15-18 août, pp. 332–336.
- Sammouri, W., Come, E., Oukhellou, L., Aknin, P., Fonlladosa, C. E. & Prendergast, K. (2012), Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework, in 'Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems', ITSC, Anchorage, AK, 16-19 sept., pp. 1351–1356.
- SAS Institute Inc, . (2013), *Data Mining Using SAS ® Enterprise Miner TM: A Case Study Approach*, third edn, Cary, NC, USA.
- Savasere, A., Omiecinski, E. & Navathe, S. B. (1995), An efficient algorithm for mining association rules in large databases, in U. Dayal, P. M. D. Gray & S. Nishio, eds, 'Proceedings of the 21th International Conference on Very Large Data Bases', VLDB '95, Morgan Kaufmann Publishers Inc., Zurich, Suisse, 11-15 sept., pp. 432–444.
- Schreiber, G. (2008), Chapter 25 knowledge engineering, in F. van Harmelen, V. Lifschitz & B. Porter, eds, 'Handbook of Knowledge Representation', Vol. 3 of *Foundations of Artificial Intelligence*, Elsevier, pp. 929–946.
- Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H. & Van de Velde, W. (1994), 'Commonkads: A comprehensive methodology for kbs development', *IEEE Expert - Intelligent Systems & their Applications* **9**(6), 28–37.
- Sebag, M. & Schoenauer, M. (1988), Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases, in J. H. Boose, B. R. Gaines & M. Linster, eds, 'Proceedings of the European Knowledge Acquisition Workshop', EKAW'88, Gesellschaft für Mathematik und Datenverarbeitung mbH, 19-23 juin, pp. 28–1 – 28–20.
- Shahbaz, M. S., Harding, J. A. & Turner, M. (2006), Product design and manufacturing process improvement using association rules, in 'Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture', Vol. 220, Professional Engineering Publishing, pp. 243–254.
- Shen, L., Tay, F. E. H., Qu, L. & Shen, Y. (2000), 'Fault diagnosis using rough sets theory', *Computers in Industry* **43**(1), 61–72.
- Silberschatz, A. (1995), On subjective measures of interestingness in knowledge discovery, in U. M. Fayyad & R. Uthurusamy, eds, 'Proceedings of the First International Conference on Knowledge

- Discovery and Data Mining', KDD-95, AAAI Press, Montréal, Québec, Canada, 20-21 août, pp. 275–281.
- Silberschatz, A. & Tuzhilin, A. (1996a), User-assisted knowledge discovery: How much should the user be involved, in 'Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery', DMKD'96, Montréal, Canada, 2 juin.
- Silberschatz, A. & Tuzhilin, A. (1996b), 'What makes patterns interesting in knowledge discovery systems', *IEEE Transactions on Knowledge and Data Engineering* **8**(6), 970–974.
- Sowa, J. F. (1984), *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Sowa, J. F. (2000), *Knowledge representation: logical, philosophical, and computational foundations*, Vol. 13, MIT Press, Pacific Grove, CA, USA.
- Sowa, J. F. & Zachman, J. A. (1992), 'Extending and formalizing the framework for information systems architecture', *IBM Systems Journal* **31**(3), 590–616.
- Staples, D. S., Greenaway, K. & McKeen, J. D. (2001), 'Opportunities for research about managing the knowledge-based enterprise', *International Journal of Management Reviews* **3**(1), 1–20.
- Stewart, T. A. (1997), 'Intellectual capital: The new wealth of organizations', Doubleday/Currency, New York, NY, USA.
- Stuber, A. (2007), Co-construction de sens par négociation pour la réutilisation en situation de l'expérience tracée, Thèse de doctorat, Université Lyon 1, France.
- Studer, R., Benjamins, V. R. & Fensel, D. (1998), 'Knowledge engineering: Principles and methods', *Data & Knowledge Engineering* **25**(1-2), 161–197.
- Sun, Z. (2004), A waterfall model for knowledge management and experience management, in M. Ishikawa, S. Hashimoto, M. Paprzycki, E. Barakova, K. Yoshida, M. Koppen, D. W. Corne & A. Abraham, eds, 'Proceedings of the 4th International Conference on Hybrid Intelligent Systems', HIS'04, IEEE, Kitakyushu, Japon, 5-8 déc., pp. 472–475.
- Swanson, L. (2003), 'An information-processing model of maintenance management', *International Journal of Production Economics* **83**(1), 45–64.
- Thomopoulos, R., Bourguet, J. R., Cuq, B. & Ndiaye, A. (2010), 'Answering queries that may have results in the future: A case study in food science', *Knowledge-Based Systems* **23**(5), 491–495.
- Tisseyre, R. C. (1999), *Knowledge management : théorie et pratique de la gestion des connaissances*, Hermes Science Publications.
- Tsai, C. Y., Chang, P. C. & Wang, S. J. (2003), 'Applying association-rule techniques and artificial neural networks to product development', *Journal of the Chinese Institute of Industrial Engineers* **20**(2), 101–112.
- Tseng, M.-C., Lin, W.-Y. & Jeng, R. (2007), Incremental maintenance of ontology-exploiting association rules, in 'International Conference on Machine Learning and Cybernetics', Vol. 4, IEEE, Hong Kong, China, 19-22 août, pp. 2280–2285.
- Uschold, M. & Grüninger, M. (1996), 'Ontologies: Principles, methods and applications', *The Knowledge Engineering Review* **11**(2), 93–136.

- 
- Villeneuve, E. (2012), Hybridation des retours d'expérience statistique et cognitif pour l'évaluation des risques: application à la déconstruction des aéronefs, Thèse de doctorat, Institut National Polytechnique de Toulouse (INPT), France.
- Vogel, C. (1988), *Génie cognitif*, Vol. 1 of *Sciences cognitives*, Masson, Paris.
- Wang, G. & Wang, Y. (2009), '3dm: Domain-oriented data-driven data mining', *Fundamenta Informaticae* **90**(4), 395–426.
- Wang, H. & Wang, S. (2008), 'A knowledge management approach to data mining process for business intelligence', *Industrial Management & Data Systems* **108**(5), 622–634.
- Wang, K. (2007), 'Applying data mining to manufacturing: the nature and implications', *Journal of Intelligent Manufacturing* **18**(4), 487–495.
- Weber, R. O. & Aha, D. W. (2003), 'Intelligent delivery of military lessons learned', *Decision Support Systems* **34**(3), 287–304.
- Weber, R. O., Aha, D. W. & Becerra-Fernandez, I. (2001), 'Intelligent lessons learned systems', *Expert Systems with Applications* **20**(1), 17–34.
- Weka (2003), *Guide Pratique du Responsable Maintenance*, Tome 2, éditions Weka.
- Wu, X., Yu, P. S., Piatetsky-Shapiro, G., Cercone, N., Lin, T. Y., Kotagiri, R. & Wah, B. W. (2003), 'Data mining: How research meets practical development?', *Knowledge and Information Systems* **5**(2), 248–261.
- Yao, H. & Etzkorn, L. (2006), 'Automated conversion between different knowledge representation formats', *Knowledge-Based Systems* **19**(6), 404–412.
- Young, T., Fehskens, M., Pujara, P., Burger, M. & Edwards, G. (2010), Utilizing data mining to influence maintenance actions, in 'Proceedings AUTOTESTCON Systems Readiness Technology Conference', IEEE, Orlando, FL, USA, 13-16 sept., pp. 267–271.
- Yu, K. M. & Zhou, J. (2010), 'Parallel tid-based frequent pattern mining algorithm on a pc cluster and grid computing system', *Expert Systems with Applications* **37**(3), 2486–2494.
- Zaki, M. J. (2000), Generating non-redundant association rules, in R. Ramakrishnan, S. Stolfo, R. Bayardo & I. Parsa, eds, 'Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '00, Boston, MA, USA, 20-23 août, pp. 34–43.
- Zaki, M. J. & Hsiao, C. J. (2002), Charm: An efficient algorithm for closed itemset mining, in R. Grossman, J. Han, V. Kumar, H. Mannila & R. Motwani, eds, 'Proceedings of the Second SIAM International Conference on Data Mining', Arlington, VA, USA, 11-13 avril, pp. 457–473.
- Zaki, M. J. & Ogihara, M. (1998), Theoretical foundations of association rules, in 'Proceedings of the 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery', Seattle, Washington, USA, 2-4 juin, pp. 1–8.
- Zhang, R. Q. & Yang, J. L. (2006), Association rules based research on man-made mistakes in aviation maintenance: A case study, in Y. Chen & A. Abraham, eds, 'Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications', Vol. 1 of *ISDA'06*, IEEE, Jinan, Chine, 16-18 oct., pp. 545–550.

Zighed, D. A. & Rakotomalala, R. (2002), Extraction de connaissances à partir de données (ecd), *in* 'Bases de données', Référence H3744, Editions T.I.



## BIBLIOGRAPHIE DE L'AUTEUR

- Potes Ruiz, P. A., Kamsu Foguem, B. & Grabot, B. (2012), Modélisation des connaissances par les graphes conceptuels pour la maintenance corrective, *in* '16es journées STP du GdR MACS', 29-30 mars, Albi, France.
- Potes Ruiz, P. A., Kamsu Foguem, B. & Grabot, B. (2013a), Amélioration des stratégies de maintenance par retour d'expérience, *in* 'Congrès des Doctorants 2013 EDSYS (Ecole Doctorale Systèmes)', 16 avril, Tarbes, France.
- Potes Ruiz, P. A., Kamsu Foguem, B. & Grabot, B. (2013b), Génération de connaissances pour la maintenance à partir du retour d'expérience, *in* '5èmes Journées Doctorales / Journées Nationales MACS (JD-JN-MACS)', 11-12 juillet, Strasbourg, France, pp. 109–114.
- Potes Ruiz, P. A., Kamsu-Foguem, B. & Grabot, B. (2013c), Improving maintenance strategies from experience feedback, *in* 'Proceedings of the 7th IFAC Conference on Manufacturing Modelling, Management and Control', Vol. 7 of *MIM'2013*, 19-21 juin, Saint Pétersburg, Russie, pp. 625–630.
- Potes Ruiz, P. A., Kamsu-Foguem, B. & Grabot, B. (2014a), 'Generating knowledge in maintenance from experience feedback', *Knowledge-Based Systems* **68**, 4–20.
- Potes Ruiz, P. A., Kamsu-Foguem, B. & Grabot, B. (2014b), An interactive approach for the post-processing in a KDD process, *in* 'Proceedings of the International Conference on Advances in Production Management Systems', Vol. 438 of *APMS 2014*, Springer Berlin Heidelberg, 20-24 septembre, Ajaccio, France, pp. 93–100.
- Potes Ruiz, P. A., Kamsu-Foguem, B. & Noyes, D. (2013), 'Knowledge reuse integrating the collaboration from experts in industrial maintenance management', *Knowledge-Based Systems* **50**, 171–186.
- Potes Ruiz, P. A., Noyes, D. & Kamsu-Foguem, B. (2012), Raisonement collaboratif à partir de cas dans la résolution de problèmes en maintenance, *in* 'Proceedings of the 9th International Conference on Modeling, Optimization & SIMulation', MOSIM'12, 6-8 juin, Bordeaux, France.





# ANNEXES

## Annexe 1. Mesures d'intérêt associées aux règles (Azé, 2003)

Mesure	Expression	Référence
Support	$P(AB)$	(Agrawal et al., 1993)
Confiance	$P(B A)$	(Agrawal et al., 1993)
Rappel	$P(A B)$	(Lavrac et al., 1999)
Confiance centrée	$P(B A) - P(B)$	(Lallich & Teytaud, 2004)
Lift	$\frac{P(AB)}{P(A)P(B)}$	(International Business Machines, 1996)
Moindre contradiction	$\frac{P(AB) - P(A\bar{B})}{P(B)}$	(Azé & Kodratoff, 2002)
Piatetsky-Shapiro	$nP(A)(P(B A) - P(B))$	(Piatetsky-Shapiro, 1991)
Lœvinger	$\frac{P(B A) - P(B)}{P(\bar{B})}$	(Lœvinger, 1947)
Corrélation	$\frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(A)P(B)P(B)}}$	(Lerman, 1981)
Indice d'implication	$q_3(A, \bar{B}) = \sqrt{n} \frac{P(A\bar{B}) - P(A)P(\bar{B})}{\sqrt{P(A)P(\bar{B})}}$	(Lerman et al., 1981)
Pearl	$P(A) *  P(B A) - P(B) $	(Pearl, 1988)
Nouveauté	$P(AB) - P(A)P(B)$	(Lavrac et al., 1999)
Satisfaction	$\frac{P(\bar{B}) - P(\bar{B} A)}{P(\bar{B})}$	(Lavrac et al., 1999)
Spécificité	$P(\bar{A} \bar{B})$	(Lavrac et al., 1999)
Fiabilité négative	$P(\bar{B} \bar{A})$	(Lavrac et al., 1999)
J-mesure	$P(AB) \log \frac{P(AB)}{P(A)P(B)} + P(A\bar{B}) \log \frac{P(A\bar{B})}{P(A)P(\bar{B})}$	(Goodman & Smyth, 1988)
Sebag-Schœnauer	$\frac{P(AB)}{P(A\bar{B})}$	(Sebag & Schoenauer, 1988)
Conviction	$\frac{P(A)P(\bar{B})}{P(A\bar{B})}$	(Brin, Motwani & Silverstein, 1997)
Intensité d'implication	$\varphi(A, B) = 1 - \Phi(q_3(A, \bar{B}))$	(Gras, 1979)
Intensité d'implication entropique	$\Psi(A, B) = \sqrt{\varphi(A, B) * \tau(A, B)}$	(Gras et al., 2001)

**Annexe 2. Règles d'association (*minsup* = 40%, *minconf* = 100%) - Cas # 3 (Test 1)**

Règle #	Règle	Sup (%)	Conf (%)
R1	<i>Nature de l'objet</i> =Piping & fittings → <i>Groupe de planification</i> =PU1	41,7	100
R2	<i>Nature de l'objet</i> =Piping & fittings → <i>Site</i> =FRXX	41,7	100
R3	<i>Durée de la défaillance</i> =0-2h → <i>Site</i> =FRXX	42	100
R4	<i>Type d'ordre</i> =PM11 → <i>Site</i> =FRXX	70,5	100
R5	<i>Unité</i> =A01 → <i>Groupe de planification</i> =PU1	87,7	100
R6	<i>Unité</i> =A01 → <i>Site</i> =FRXX	87,7	100
R7	<i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	98,5	100
R8	<i>Nature de l'objet</i> =Piping & fittings, <i>Unité</i> =A01 → <i>Groupe de planification</i> =PU1	40,7	100
R9	<i>Nature de l'objet</i> =Piping & fittings, <i>Unité</i> =A01 → <i>Site</i> =FRXX	40,7	100
R10	<i>Nature de l'objet</i> =Piping & fittings, <i>Site</i> =FRXX → <i>Groupe de planification</i> =PU1	41,7	100
R11	<i>Nature de l'objet</i> =Piping & fittings, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	41,7	100
R12	<i>Nature de l'objet</i> =Piping & fittings → <i>Site</i> =FRXX, <i>Groupe de planification</i> =PU1	41,7	100
R13	<i>Durée de la défaillance</i> =0-2h, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	41,1	100
R14	<i>Type d'ordre</i> =PM11, <i>Unité</i> =A01 → <i>Groupe de planification</i> =PU1	62,9	100
R15	<i>Type d'ordre</i> =PM11, <i>Unité</i> =A01 → <i>Site</i> =FRXX	62,9	100
R16	<i>Type d'ordre</i> =PM11, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	69,4	100
R17	<i>Unité</i> =A01, <i>Site</i> =FRXX → <i>Groupe de planification</i> =PU1	87,7	100
R18	<i>Unité</i> =A01, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	87,7	100
R19	<i>Unité</i> =A01 → <i>Site</i> =FRXX, <i>Groupe de planification</i> =PU1	87,7	100
R20	<i>Nature de l'objet</i> =Piping & fittings, <i>Unité</i> =A01, <i>Site</i> =FRXX → <i>Groupe de planification</i> =PU1	40,7	100
R21	<i>Nature de l'objet</i> =Piping & fittings, <i>Unité</i> =A01, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	40,7	100
R22	<i>Nature de l'objet</i> =Piping & fittings, <i>Unité</i> =A01 → <i>Site</i> =FRXX, <i>Groupe de planification</i> =PU1	40,7	100
R23	<i>Type d'ordre</i> =PM11, <i>Unité</i> =A01, <i>Site</i> =FRXX → <i>Groupe de planification</i> =PU1	62,9	100
R24	<i>Type d'ordre</i> =PM11, <i>Unité</i> =A01, <i>Groupe de planification</i> =PU1 → <i>Site</i> =FRXX	62,9	100
R25	<i>Type d'ordre</i> =PM11, <i>Unité</i> =A01 → <i>Site</i> =FRXX, <i>Groupe de planification</i> =PU1	62,9	100

Annexe 3. Règles d'association (*minsup* = 10%, *minconf* = 70%) - Cas # 3 (Test 2)

Règle #	Règle	Sup (%)	Conf (%)
R1	Groupe d'équipement=DIV → Criticité de l'équipement=#N/A	11,5	100
R2	Description de l'objet=#N/A → Date de création=sept-oct-nov-dec	12,1	98,4
R3	Groupe d'équipement=100 → Criticité de l'équipement=Medium	13,4	83,7
R4	Effet de la défaillance=Vide → Nature de l'objet=Vide	13,5	76,8
R5	Nature de l'objet=Vide → Effet de la défaillance=Vide	13,5	70,1
R6	Effet de la défaillance=Vide → Type de cause=Vide	14,8	84,2
R7	Type de cause=Vide → Nature de l'objet=Vide	16	71,6
R8	Nature de l'objet=Vide → Type de cause=Vide	16	82,6
R9	Groupe d'équipement=500 → Criticité de l'équipement=High	16,3	73,9
R10	Type de cause=Wear and tear → Nature de l'objet=Piping & fittings	19,9	77,5
R11	Effet de la défaillance=Leak → Nature de l'objet=Piping & fittings	23,2	75,7
R12	Effet de la défaillance=Vide, Type de cause=Vide → Nature de l'objet=Vide	13,2	88,7
R13	Effet de la défaillance=Vide, Nature de l'objet=Vide → Type de cause=Vide	13,2	97,2
R14	Nature de l'objet=Vide, Type de cause=Vide → Effet de la défaillance=Vide	13,2	82,5
R15	Effet de la défaillance=Vide → Nature de l'objet=Vide, Type de cause=Vide	13,2	74,7
R16	Effet de la défaillance=Vide, Durée de la défaillance=0 → Type de cause=Vide	10,2	90,1
R17	Type de cause=Vide, Durée de la défaillance=0 → Effet de la défaillance=Vide	10,2	71,4
R18	Type de cause=Vide, Durée de la défaillance=0 → Nature de l'objet=Vide	10,9	76,6
R19	Nature de l'objet=Vide, Durée de la défaillance=0 → Type de cause=Vide	10,9	90,7
R20	Type de cause=Wear and tear, Effet de la défaillance=Leak → Nature de l'objet=Piping & fittings	13,5	86,9
R21	Type de cause=Wear and tear, Durée de la défaillance=0-2h → Nature de l'objet=Piping & fittings	10,8	82,8
R22	Groupe d'équipement=400, Nature de l'objet=Piping & fittings → Criticité de l'équipement=High	11,1	78,9
R23	Effet de la défaillance=Leak, Criticité de l'équipement=Medium → Nature de l'objet=Piping & fittings	11,3	82,4
R24	Effet de la défaillance=Leak, Criticité de l'équipement=High → Nature de l'objet=Piping & fittings	10	71
R25	Effet de la défaillance=Leak, Coût total réel=0<x<200 → Nature de l'objet=Piping & fittings	12,4	87
R26	Effet de la défaillance=Leak, Durée de la défaillance=0-2h → Nature de l'objet=Piping & fittings	12,4	83,7





## **Génération de connaissances à l'aide du retour d'expérience : Application à la maintenance industrielle**

Les progrès des technologies de l'information et de la communication et la généralisation des systèmes d'information tels que les ERP (*Enterprise Resource Planning*) en milieu industriel rendent disponible un grand volume d'informations issues des événements ou des faits passés, dont la réutilisation devient un enjeu majeur pour les entreprises.

Nous proposons dans ce travail une démarche originale permettant de générer de nouvelles connaissances à partir des expériences passées des entreprises, en combinant deux approches émergentes complémentaires : la démarche de Retour d'Expérience (REx) et les techniques d'Extraction de Connaissances à partir de Données (ECD). Dans cette démarche, nous proposons également une méthodologie structurée pour interpréter et évaluer l'intérêt des connaissances extraites par les experts du domaine industriel. Cette démarche est validée sur des cas réels portant sur la maintenance industrielle.

**Mots clés :** Retour d'expérience, Extraction de connaissances à partir de données, Formalisation des connaissances, Extraction de règles d'association, Maintenance industrielle

## **Knowledge generation using experience feedback: Application to industrial maintenance**

The recent advances in information and communication technologies and the generalisation of information systems such as ERPs (*Enterprise Resource Planning*) in industrial environments make available a large amount of information related to past events or historical facts. The reuse of this information is becoming a major issue for companies.

We suggest in this work an original approach allowing to generate new knowledge based on previous experiences in companies, by combining two emerging and complementary approaches: the Experience Feedback (EF) process and the Knowledge Discovery techniques from Databases (KDD). In this approach, we also suggest a structured methodology aiming at the interpretation and evaluation of the usefulness of the extracted knowledge by industrial domain experts. This approach is validated on real cases dealing with the industrial maintenance.

**Keywords:** Experience feedback, Knowledge discovery from databases, Knowledge formalization, Association rule mining, Industrial maintenance.